The Molecular Biology of Sickle Cell Anaemia

Shannon, Matthew Frederick

Awarding institution:
King's College London

THE MOLECULAR BIOLOGY OF SICKLE CELL ANAEMIA

Matthew F. Shannon

1227783

Department of Medical & Molecular Genetics

June 2017

Thesis submitted to King's College London in fulfilment of the degree of

Doctor of Philosophy

Declaration

I hereby declare that the work presented in this PhD thesis is my own, with the exception of the contributions stated here.

Processing of DNA samples for Whole Exome Sequencing library preparation and sequencing was carried out by Simon Hazelwood-Smith or Dr Ines Barbosa, and sequencing was performed by the NIHR Biomedical Research Centre Genomics Core Facility at Guy's and St Thomas' NHS Foundation Trust. For experiments performed using Sickle Cell Anaemia patient blood samples, suitable patients were identified and recruited by clinical collaborators Professor Swee Lay Thein and Dr Catherine Gardner at King's College Hospital NHS Foundation Trust, and Dr Jo Howard at Guy's and St Thomas' NHS Foundation Trust. All Fluorescence-Activated Cell Sorting (FACS) was performed by the BRC Flow Cytometry Core Facility at Guy's Hospital. Additionally, Prodromos Chatzikyriakou assisted with the initial cloning steps in construction of the CRISPR-Cas9 plasmids, specifically the steps described in section 2.4.2.1.

Acknowledgements

I would like to thank my primary supervisor, Rebecca Oakey for her enormous contributions in guiding the direction of this project, for continually motivating me, and for making my time here at King's so enjoyable. I am extremely grateful to all members of the Oakey laboratory group, past and present, for training me in the experimental techniques, and providing insightful advice throughout the duration of the project, with particular thanks to Siobhan Hughes for guidance in optimising cloning techniques. I would also like to thank Reiner Schulz for his informed input, particularly regarding the bioinformatics work.

I would like to thank my secondary supervisor, Swee Lay Thein for initially proposing this project, and for the enormous help that her clinical expertise in sickle cell anaemia has provided. All the members of Swee Lay's laboratory group were incredibly helpful, and I am particularly grateful to Amandine Breton, for all her help with the cell work, and for the time spent providing valuable feedback on this thesis.

Any study involving patient participation requires a lot of clinical work, and for that I am very grateful to Catherine Gardner and Jo Howard who co-ordinated recruitment at KCH and Guy's Hospital respectively. I would also like to thank the research nurses, especially Marlene Allman, and of course all of the sickle cell anaemia patients who so generously participated, and without whom this work would have been impossible.

I am incredibly grateful for the generous funding provided by the Guy's and St Thomas' Charity Prize PhD Programme that enabled me to undertake my studies at KCL, and to the King's Health Partners Research and Development Challenge Fund for further funding this work. I would also like to thank the teams at the BRC Flow Cytometry and Genomics core facilities, who were always extremely helpful. I am also grateful to Michael Simpson for his computational support.

I would like to thank Christos, Ines, Jake, Laura and Seth, as well as many others for the nights of pool and sambuca at Guy's Club, and for generally keeping me sane. I would also like to offer special thanks to Jackie for her support and optimism throughout this very stressful process.

Finally, I would like to thank my parents, George and Caroline Shannon, for their continued encouragement over the years, and my siblings, Nicholas Gregory, Jemima, and Edward, for many heated discussions verging closer to science fiction than actual science.

Abstract

Sickle cell anaemia (SCA) is a haemolytic anaemia that reduces life expectancy and places a great burden on healthcare systems worldwide. Despite being a monogenic disorder, the phenotypic severity varies greatly between patients, ranging from patients that experience multiple strokes and organ failure during childhood, to those that live largely unaffected lives. Some genetic variants that affect globin gene expression are known to influence phenotype severity, but most of this variation remains unaccounted for.

We conducted whole exome sequencing analyses, comparing SCA patients with mild and severe clinical phenotypes, with the aim of identifying novel genetic modifiers of the disease. SCA patient exomes were sequenced from a cohort at King's College Hospital, and combined with publicly available SCA exomes recruited in the United States. Nine candidate variants were identified in genes with plausible mechanisms to influence the pathophysiology of the disease. The genes identified in this study affected nitric oxide signalling, haematopoietic regulation, globin gene expression and recovery from ischaemic injury.

In order to evaluate these variants, a CRISPR genomic editing pipeline was established and tested on two previously identified candidate modifiers of SCA, in the genes ASH1L and KLF1. These variants were successfully introduced into erythroleukaemic cells and provide a pathway for testing the novel modifier genes identified in the exome sequencing analysis. Preliminary studies indicate that both ASH1L and KLF1 variants alter globin gene expression.

In addition to genetic factors, we also hypothesised that epigenetic factors affect the SCA phenotype, and play a role in the therapeutic mechanism of hydroxyurea treatment. We optimised a method for isolating $CD45^{+}CD71^{+}GPA^{-}$ nucleated erythroid progenitors from small volumes of SCA peripheral blood. This was undertaken to evaluate the role of the epigenome in SCA phenotype severity and drug action, but for which patient sample collection proved too challenging within our clinical cohort.

Abbreviations

3C	Chromosome Conformation Capture
BPG	2,3-bisphosphoglycerate
CADD	Combined Annotation Dependant Depletion
Cascade	CRISPR-Associated Complex for Antiviral Defence
ChIP	Chromatin Immunoprecipitation
CLP	Common Lymphoid Progenitor
CMP	Common Myeloid Progenitor
CRISPR	Clustered Regularly Interspaced Palindromic Repeats
crRNA	CRISPR RNA
DSB	Double Strand Break
EPO	Erythropoietin
EPOR	Erythropoietin Receptor
FACS	Fluorescence-Activated Cell Sorting
FLAGS	Frequently mutated Genes
gDNA	Genomic DNA
GMP	Granulocyte-Macrophage Progenitor
GPA	Glycophorin A
HAT	Histone Acetyltransferase
HbA	Adult Haemoglobin
HbF	Foetal Haemoglobin
HbS	Sickle Haemoglobin
HDAC	Histone Deacetylase
HDR	Homology Directed Repair
HPFH	Hereditary Persistence of Foetal Haemoglobin
HPLC	High Performance Liquid Chromatography
HSCs	Haematopoietic Stem Cells
HU	Hydroxyurea
IGF1	Insulin-like Growth Factor 1
IL-3	Interleukin-3

LCR	Locus Control Region
LINC	Long Intergenic ncRNA
MEP	Megakaryocyte-Erythroid Progenitor
MPNs	Myeloproliferative Neoplasms
MPP	Multipotent Progenitor
MYDGF	Myeloid Derived Growth Factor
NETs	Neutrophil Extracellular Traps
NHEJ	Non-Homologous End Joining
NLS	Nuclear Localisation Signal
NO	Nitric Oxide
PAM	Protospacer Adjacent Motif
PBMCs	Peripheral Blood Mononuclear Cells
PBS	Phosphate-Buffered Saline
PCR	Polymerase Chain Reaction
RFLPs	Restriction Fragment Length Polymorphisms
RNR	Ribonucleotide Reductase
SCA	Sickle Cell Anaemia
SCD	Sickle Cell Disease
SCF	Stem Cell Factor
SDM	Site Directed Mutagenesis
sgRNA	Short Guide RNA
snoRNA	Small Nucleolar RNA
SNP	Single Nucleotide Polymorphism
ssODN	Single Stranded Oligodeoxynucleotides
TCD	Transcranial Doppler Ultrasonography
tracrRNA	Trans-Activating CRISPR RNA
UTR	Untranslated Region
WES	Whole Exome Sequencing
ZFN	Zinc Finger Nuclease

Table of Contents

Appendices 7 - 12 are provided as additional data files, and are not printed in this volume:

Appendix 7: List of 4988 ncRNA that were found to have variants annotated in both the SureSelect and NimbleGen exome groups. This list was generated using variants annotated to genome build GRCh37/hg19.

Appendix 8: List of the 2,442 protein coding variants identified by the Fisher's Exact Tests in 4.5.2, that were found to be significantly enriched in either the Mild or Severe patient groups. Variants are annotated to genome build GRCh37/hg19, and the homozygous and heterozygous counts are given for both the Mild and Severe groups.

Appendix 9: List of the 236 protein coding variants identified by the Fisher's Exact Tests in 4.5.3, that were found to be significantly enriched in either the SWiTCH or HUSTLE patient groups. Variants are annotated to genome build GRCh37/hg19, and the homozygous and heterozygous counts are given for both the SWiTCH and HUSTLE groups.

Appendix 10: List of 2,556 genes that were identified as being commonly mutated in exome sequencing studies. Genes were included based on their presence in lists published by Fuentes *et al.* (2012) or Shyr *et al*. (2015), as described in detail in 4.3.1.7.

Appendix 11: List of 7,420 genes and transcripts found not to be expressed in Fantom5 expression data for human haematopoietic tissues, as described in 4.3.1.8. The Fantom5 datasets used to generate this list are given in Appendix 6. Genes were considered to be haematopoietically silent if they had an RNA expression level of <1tpm (transcripts per million) in all of the datasets analysed.

Appendix 12: Full list of the 11,419 variants from the Mild SCA patient exomes, that were not excluded using the variant filtering pipeline developed in this project, as described in 4.3.1. Variants are annotated to genome build GRCh37/hg19, and the homozygous and heterozygous counts are given for the Mild, Severe, SWiTCH and TWiTCH groups.

Table of Figures

Table of Tables

Chapter 1 Introduction

1.1 General Introduction

This work aimed to investigate factors that affect the severity of the clinical phenotype presented by sickle cell anaemia (SCA) patients. Despite being well characterised as a monogenic disorder, the severity of symptoms, as well as the response to current treatments, varies greatly between SCA patients.

This aim was investigated through three distinct research objectives:

1. To optimise a non-invasive technique to isolate nucleated erythroid progenitors from the peripheral blood of SCA patients. There is a growing body of evidence to support the role of epigenetic factors in the pathology of SCA, both in terms of naturally occurring variation and in response to drug treatment. Due to the nature of erythrocytes and their lack of a nucleus, it is notoriously difficult to investigate epigenetic regulation in these cells *in vivo*, and most existing studies either focus on transcriptomic analyses of enucleated cells, or use *in vitro* treatment models. We set out to develop a protocol to allow investigation of both the epigenome & transcriptome of a nucleated erythrocyte progenitor population in a longitudinal manner (i.e. detecting changes in response to drug treatment *in vivo*).

2. To conduct a whole exome sequencing (WES) study, comparing sequencing data from SCA patients with severe and mild clinical phenotypes. At one end of the phenotypic spectrum, some patients experience very few symptoms and live largely unaffected lives, while at the other end of the spectrum, some patients experience multiple strokes and organ damage at young age, and are frequently hospitalised. It has been shown that sequence variation in some genes (genetic modifiers) heavily influence the pathophysiology of the disease, and are known to affect symptomatic severity. However, much of this variation remains unaccounted for. We proposed to conduct a WES study investigating the extreme ends of the phenotypic spectrum, in order to

identify novel genetic modifiers that may be influencing the severity of symptoms in these patients.

3. To use CRISPR genomic editing to replicate two previously identified candidate modifier SNPs *in vitro*. The aim of this work was to perform preliminary functional analyses to inform on the effect these SNPs have on gene function, as well as to set up a CRISPR genomic editing pipeline in our laboratory, to allow functional analyses of candidate variants identified by the WES study in the future.

1.2 Haemoglobin & SCA

1.2.1 Healthy Haemoglobin

Haemoglobin is a tetrameric protein expressed at high levels in erythrocytes, with the ability to bind oxygen molecules (O_2) under conditions of high O_2 concentration, and release them under conditions of low O_2 concentration, allowing it to efficiently distribute O_2 to tissues that need it. The affinity of haemoglobin for O_2 binding is allosterically regulated by multiple small molecules, including 2,3-bisphosphoglycerate (BPG), Cl^- and H^+, allowing for tight regulation over the O_2 distribution process[1–5]. This is additionally regulated throughout development by the use of various isoforms of haemoglobin, controlled through the expression of different subunits. Since in the foetus O_2 must be sourced from the mother's blood through the placenta, where O_2 availability is much lower than in the lungs, a higher affinity for O_2 binding is required for foetal/embryonic haemoglobin. This ensures that at the low concentration at which O_2 dissociates from the maternal haemoglobin, it still binds to the foetal/embryonic haemoglobin for transport through the foetal/embryonic body[6–8].

The haemoglobin tetramer is made up of two α-globin like subunits and two β-globin like subunits, which are encoded by two distinct gene clusters on chromosomes 16 and 11, respectively. The layout of the genes at these loci are shown in Figure 1.1, and the genes are positioned in the order in which they are expressed throughout development[9]. Two Gower haemoglobins are expressed during embryonic stages, HbGower I ($\zeta_2\varepsilon_2$) and HbGower II ($\alpha_2\varepsilon_2$), and during foetal development expression of ζ-globin is completely replaced by the two α-globin genes, and ε-globin is replaced by expression of the γ-globin genes, giving rise to the foetal haemoglobin (HbF: $\alpha_2\gamma_2$)[9,10]. Shortly after birth, expression is switched from the γ-globins to β-globin and δ-globin, producing adult haemoglobins HbA ($\alpha_2\beta_2$) and HbA2 ($\alpha_2\delta_2$). HbA is the most abundant form of haemoglobin in healthy adults, making up >95% of total haemoglobin[11,12].

Each globin subunit folds around an aqueous pocket containing a protoporphyrin IX molecule, which has a negatively charged central ring that coordinates the binding of an Fe^{2+} ion[13]. Two histidine residues from the globin peptide interact with Fe^{2+}, the first is situated below the plane of the porphyrin ring and is referred to as the proximal histidine, interacting directly with Fe^{2+}, and the second is situated above the plane of the porphyrin ring, interacting with Fe^{2+} through coordination of a bound O_2 molecule, and is referred to as the distal histidine[14,15].

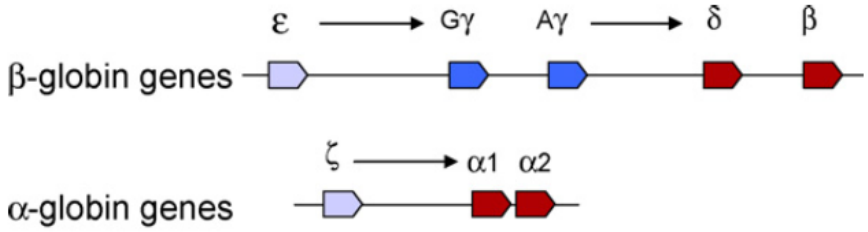


Figure 1.1: Layout of the α-globin like gene locus on chromosome 16 and the β-globin like gene locus on chromosome 11. Genes are positioned in the order in which they are expressed during development. Embryonic haemoglobin – $\zeta_2\varepsilon_2$, foetal haemoglobin – $\alpha_2\gamma_2$, adult haemoglobin $\alpha_2\beta_2$. Adapted from Kiefer *et al.* 2008[9].

In the absence of O_2 binding, Fe^{2+} sits just below the plane of the porphyrin ring, closer to the proximal histidine[13]. Upon O_2 binding, electron rearrangements reduce the size of Fe^{2+} and it is pulled into the centre of the porphyrin ring, forming a more stable structure coordinated by six interactions[13,16]. This is illustrated in Figure 1.2.

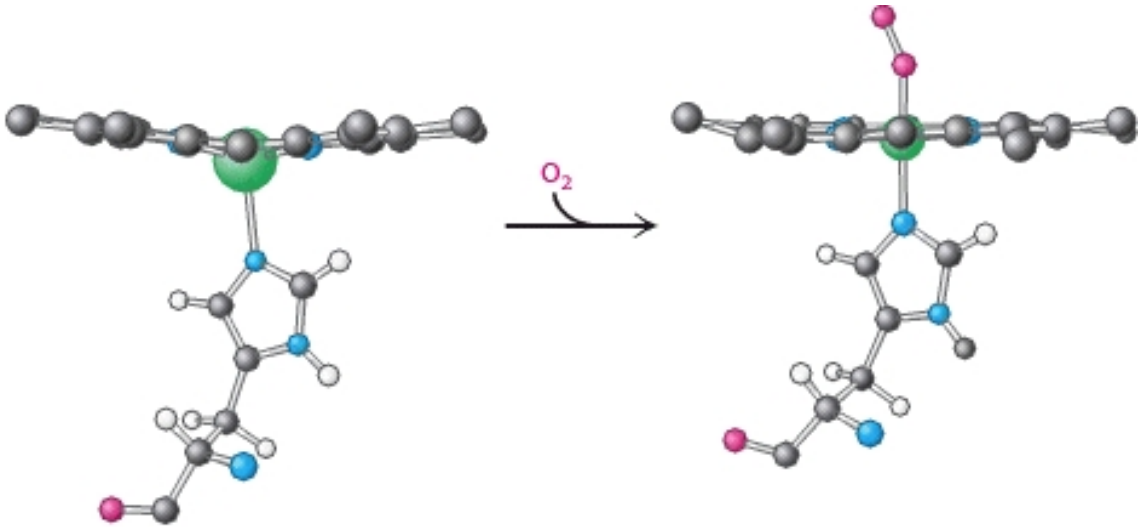


Figure 1.2: Oxygen binding stabilises the coordination of Fe^{2+} (Green) in the plane of the porphyrin ring, dragging the proximal histidine (and therefore the helix to which it is attached) closer, altering the structure of the globin subunit. Image is from Berg, Tymoczko & Stryer (2002)[13].

The conformational changes that occur upon binding of O_2 to the globin subunits transition haemoglobin from the T (Tense, deoxygenated) state to the R (Relaxed, oxygenated state)[2,13,15,17]. The conformational changes in each globin subunit alter the interaction with the neighbouring subunits, allowing for the cooperative binding effect of O_2 to haemoglobin. Upon O_2 binding to one subunit, conformational changes make the R state of the other subunits more favourable, increasing O_2 affinity[17]. This effect is additive, and when three of the four subunits are oxygenated, the affinity of the fourth subunit for oxygen is increased 20-fold[13]. The allosteric

regulators of haemoglobin oxygen affinity act by stabilising either the T or R state, e.g. there is a central cavity between the four subunits that closes upon transitioning to the R state, BPG binds this cavity, preventing the conformation changes required to transition to the R-state and therefore favouring the low oxygen affinity T-state[13,18].

1.2.2 Sickle Haemoglobin (HbS) & SCA Pathophysiology

SCA is a recessive disorder caused by an aberrant haemoglobin variant, referred to as sickle haemoglobin (HbS). HbS is the result of a single nucleotide polymorphism (SNP) in the β-globin gene (HBB) on chromosome 11. The SNP is an A to T substitution, replacing the negatively charged glutamic acid at position 6 with a hydrophobic valine residue (E6V, OMIM: 141900.0243). This substitution alters the conformation of haemoglobin; in the absence of Glu-6 no intramolecular ionic interaction occurs with Lys-132, instead Val-6 forms intermolecular hydrophobic interaction with Phe-85 & Leu-88 of a neighbouring tetramer, resulting in HbS polymerisation[19,20].

Patients with SCA inherit two copies of this aberrant β-globin allele (β^S), resulting in the production of sickle haemoglobin (HbS, $\alpha_2\beta^S_2$) rather than the wild type (HbA, $\alpha_2\beta_2$). HbS has reduced solubility under low oxygen conditions, and in the deoxygenated T-state polymerises into long helical chains rather than free-floating globular tetramers[19]. These aberrant polymers aggregate and distort the shape of the erythrocyte cell membrane, forcing them from a biconcave structure to the eponymous sickle shape. Sickled erythrocytes are more rigid than wild type, and cannot pass through smaller capillaries as easily, slowing blood flow and resulting in vaso-occlusion and acute pain, commonly referred to as sickle crises[21]. The sickled cells are fragile, with an average survival time of 10-20 days compared to 110-120 days for wild type erythrocytes, leading to a chronic haemolytic anaemia[22].

High rates of haemolysis in SCA patients leads to increased levels of extra-cellular haemoglobin in the blood. The cell-free haemoglobin then binds and sequesters nitric oxide (NO), a vasodilatory signalling molecule. This reduction in NO bioavailability further increases susceptibility to frequent vaso-occlusive events and pulmonary hypertension. Recently however, the importance of the role of NO signalling in SCA pathology has been disputed, mostly due the perceived ineffectiveness of treatments designed solely to boost NO signalling[23,24].

1.2.3 Genotypes of Sickle Cell Disease

A number of other β-globin genotypes result in a variety of sickle cell phenotypes when co-inherited with at least one copy of the β^S allele. Collectively these disorders are classified as Sickle Cell Disease (SCD), with SCA referring to homozygosity for the β^S genotype.

1.2.3.1 Alternative β-globin Genotypes

The phenotype of SCD requires at least one β^S mutation, the most common genotype being HbSS (homozygosity for β^S), however there are a variety of rare alternative mutations that can give rise to the SCD phenotype when co-inherited with the β^S allele, either passively through reduction of functional β-globin levels, or by acting cooperatively[25]. Haemoglobin C is the most common of these, and is discussed below in 1.2.3.2.

1.2.3.2 Haemoglobin C

An alternative substitution at the same position as the HbS mutation (HBB Glutamic acid 6), gives rise to a milder form of the SCA phenotype. This allele is referred to as HbC, and is characterised by Glutamic acid to Lysine substitution (rs33930165), rather than the Valine substitution that is associated with HbS. Similarly to HbS, both heterozygous and homozygous forms of HbC have been associated with protection from Malaria, and as such HbC is also most prevalent in populations of African descent[26–29].

Homozygosity for HbC presents as a mild haemolytic anaemia, less severe than that observed with HbSS patients[30,31]. Similarly to that observed with HbAS, coinheritance of HbC with the wild type HbA allele results in a mostly asymptomatic phenotype[32].

Coinheritance of HbC with HbS, results in a much more severe phenotype, although still with reduced severity of some of the vasculopathy related complications observed in HbSS patients[31,33].

1.2.3.3 β-Thalassaemia

β-Thalassaemia is characterised by insufficient production of the β-globin component of adult haemoglobin, resulting in anaemia. The severity of the clinical symptoms of β-thalassaemia vary greatly, dependent on the levels of functional β-globin synthesised. Alleles that produce no functional transcripts, either due to early termination, frameshift mutations or large scale genomic deletions are classified as β^0 genotype[34–37], while alleles that contain polymorphisms

that reduce expression, typically by disrupting the promoter or other gene regulatory regions are classified as β^+ genotype[38–40]. The β^+ alleles display much more heterogeneity, based on the quantitative effect that the specific variant has on β-globin synthesis, and can be classified as severe, mild or silent[41].

Heterozygous β-thalassaemia patients ($\beta/\beta^{+/0}$) are generally asymptomatic, and are referred to as carriers for the disease. Homozygous patients ($\beta^{+/0}/\beta^{+/0}$) lack a fully functional copy of the gene, and phenotypes range from thalassaemia major, suffering severe anaemia and requiring blood transfusions from shortly after birth, to thalassaemia intermedia, and the asymptomatic state[41]. Similarly to SCA, β-thalassaemia can be ameliorated by persistent expression of γ-globin after birth (as discussed in 1.6.1), providing a functional alternative to β-globin[42–44].

In England roughly 44 per 1000 births are carriers of a β-thalassaemia allele, and coinheritance with a β^S allele is common[45]. For both SCA and β-thalassaemia, heterozygosity for the pathogenic allele is mostly asymptomatic. However, compound heterozygotes where each of the disorders is co-inherited on separate β-globin alleles gives rise to a SCD phenotype[46–48]. In these cases the severity depends on the functionality of the β-thalassaemia allele. With decreasing expression from the functional allele, the ratio of β^{wt}:β^S shifts in favour of β^S, and so levels of the pathogenic HbS increase[48–51]. Heterozygotes for β^S/β^0 produce no healthy β-globin, and present the same clinical phenotype as HbSS patients.

1.3 β-globin Locus Control

The β-globin gene locus consists of five β-globin paralogues, as shown in Figure 1.1 as well as a β-globin pseudogene (HBBP1). The genes have a highly specific pattern of expression and are arranged in the order in which they are expressed throughout development, progressing from ε-globin at the 5' end expressed in embryos to δ-globin and β-globin at the 3' end expressed in adults. Expression of the genes in the locus is developmental stage and tissue specific, and is tightly regulated by a variety of factors, including long range chromatin interaction between individual promoters and an upstream locus control region (LCR). Disruption of these regulatory mechanisms can lead to blood disorders such as anaemias and thalassaemias.

1.3.1 Transcription Factors

Transcription factors play an important role in regulating gene expression, stabilising the chromatin state surrounding the transcription start site, recruiting and stabilising the transcription initiation complex and RNA polymerases (or destabilising in the case of repressors).

The GATA family of transcription factors are key regulators of haematopoietic development[52]. GATA2 expression in haematopoietic stem cells (HSCs) is replaced by GATA1 expression at the proerythroblast stage[52]. GATA3 is also expressed at HSC stage, and is involved in the development of lymphoid lineages[53]. GATA1 is essential for globin gene expression, and knockdown in K562 cells results in chromatin reorganisation at the β-globin locus, forming transcriptionally repressive heterochromatin[54]. GATA1 upregulates expression of other erythroid transcription factors including KLF1 and TAL1, and binds with them at the β-globin locus stabilising long range chromatin interactions between CTCF/RAD21 binding sites[55–57].

TAL1 is an important haematopoietic transcription factor, critical for the establishment of haematopoietic lineages from mesodermal cells in early development, and is involved in maintenance of HSC renewal and quiescence in adults[58–60]. During haematopoiesis, expression of TAL1 is highly expressed in the myeloid lineages, and in erythroid cells associates with GATA1, LMO2 and Ldb1 at the β-globin locus, where it is required for chromatin looping[58,61,62].

Krüppel-like factor 1 (KLF1) is an erythroid specific transcription factor that plays a crucial role in the γ-globin to β-globin switch[63,64]. KLF1 directly activates β-globin expression by binding to the

promoter, and also silences transcription from the γ-globin genes through activation of BCL11A[64–67]. Interestingly, low levels of KLF1 expression are also required for ε and γ globin expression during early development, and it is thought to be required to stabilise the chromatin architecture of the β-globin locus with GATA1[57,68].

BCL11A is required for the γ-globin to β-globin switch during erythroid development, repressing expression from the γ-globin genes by recruitment of the NuRD histone deacetylase complex to the promoter regions, mediated by interactions with SOX6, GATA1 and FOG1[69–72]. Knock down of BCL11A reactivates γ-globin expression, and BCL11A polymorphisms have been associated with increased HbF levels in adults[69,73–75].

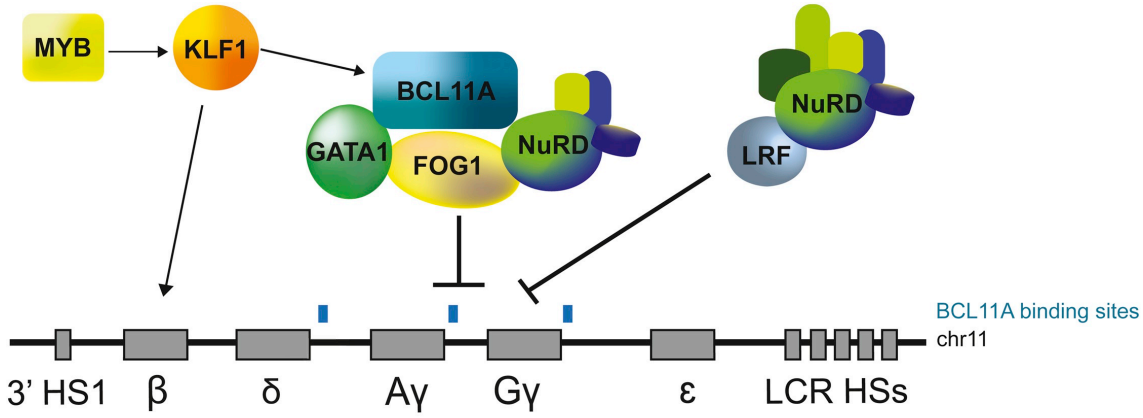


Figure 1.3: Figure illustrating the role of transcription factors during the γ-globin to β-globin switch during erythroid development. MYB activated upregulation of KLF1 causes an increase in BCL11A, as well as ZBTB7A (LRF), and both of these form complexes recruiting the NuRD repressor to the γ-globin genes. Figure adapted from Cavazzana *et al.* (2017)[76].

MYB is a haematopoietic transcription factor, required during early haematopoiesis for commitment to the erythroid lineages, and activates transcription of both KLF1 and LMO2[77,78]. MYB is activated by a distal enhancer in the MYB-HBS1L intergenic region upon binding by the transcriptional activation complex containing GATA1 and TAL1, as well as KLF1, resulting in a positive-feedback loop for transcriptional activation and commitment to the erythroid lineage[79,80]. Upregulation of KLF1 by MYB also provides a mechanism by which increased MYB promotes the γ-globin to β-globin switch. Sequence polymorphisms at the MYB-HBS1L intergenic region have been associated with increased HbF levels, and along with BCL11A and KLF1, MYB is downregulated in response to HU treatment[73,75,81].

ZBTB7A (aka LRF - Leukaemia/lymphoma-related factor) is a haematopoietic transcription factor, that is a downstream target of both GATA1 and KLF1, and plays a role in lineage determination in many haematopoietic cell populations, at both late and early stages[82–84].

ZBTB7A is also involved in the repression of the γ-globin genes through recruitment of the NuRD histone deacetylase complex, and acts independently to BCL11A[85]. Since both ZBTB7A and BCL11A are upregulated by increased KLF1 expression, this means that KLF1 dependant repression of the γ-globin genes during the globin switch occurs through two parallel pathways[82,85].

1.3.2 Chromatin Looping

Upstream of the β-globin gene locus is a cluster of 5 DNase I hypersensitivity sites (HS1-5, with HS5 being the furthest upstream), these are collectively known as the Locus Control Region (LCR), and play an important role in gene regulation. In addition to the LCR, the β-globin locus has a downstream hypersensitivity site (3'HS). HS5 and 3'HS contain CTCF binding sites in both mice and humans, and CTCF binding results in chromatin looping, bringing these two sites into close proximity, creating a chromatin domain and insulating from the effects of neighbouring enhancers (Figure 1.4A)[86,87].

Within this chromatin domain, the LCR acts as a distal enhancer, looping out into close proximity of the promoter of the specific gene being expressed (Figure 1.4B). This allows recruitment of chromatin remodelling machinery, transcriptional machinery and stabilisation of the transcription initiation complex. HS1-4 of the LCR are required for globin gene expression, and loss of HS5 does not affect expression levels, suggesting that its role is restricted to forming the chromatin domain with 3'HS[88].

The genes in both the α-globin and β-globin loci are arranged in the order in which they are expressed in development. Interestingly, the order of expression changes if the genes are rearranged, showing that the spatial organisation of the locus is important. This is thought to be due to the distance of the promoters from the LCR affecting the affinity for chromatin looping[89].

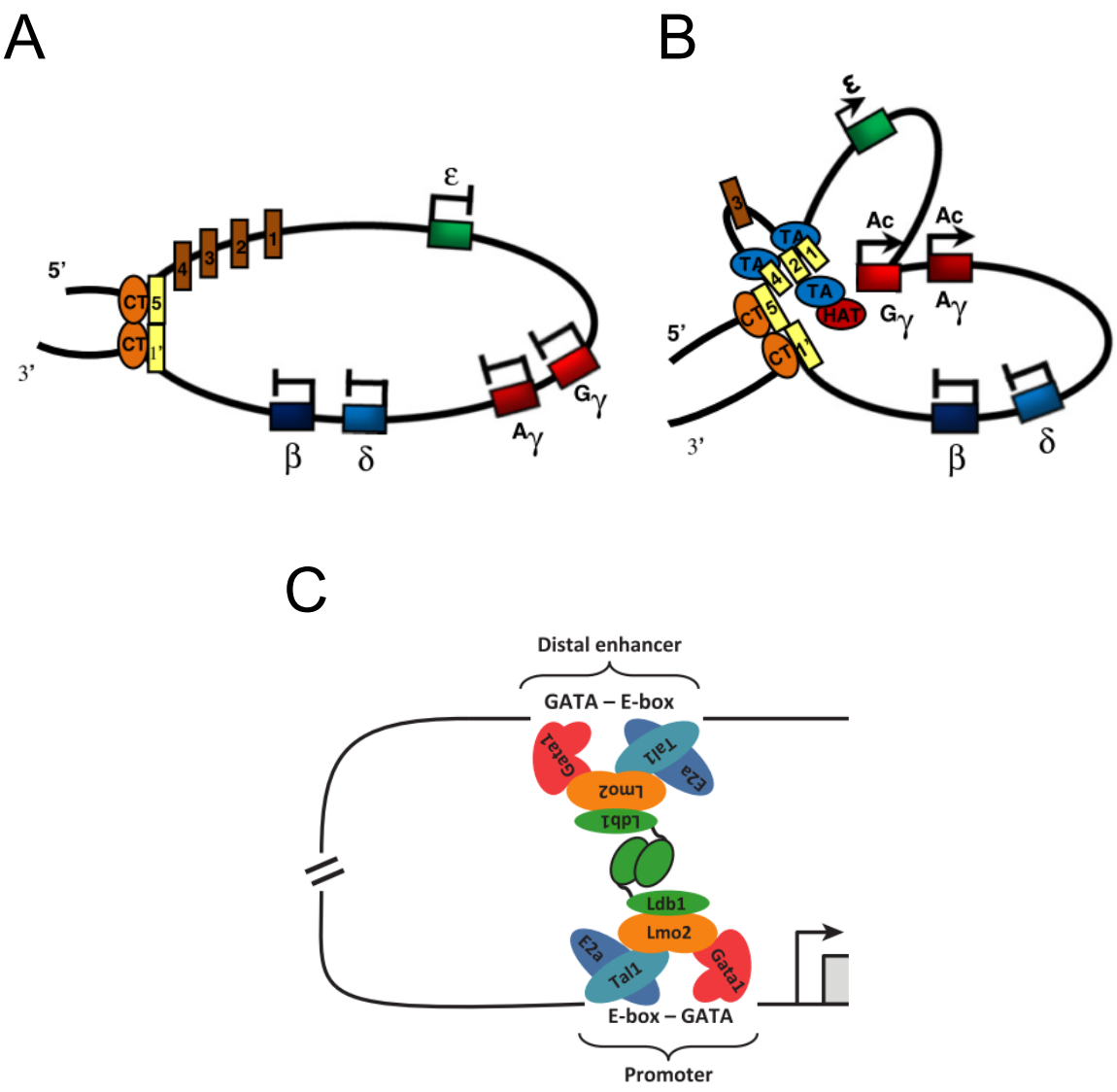


Figure 1.4: Chromatin looping at the β-globin locus. A – Interactions between CTCFs (orange) at HS5 and 3'HS (yellow) form a chromatin domain. B – Interactions between HS1, 2 & 4 and the γ-globin promoters result in histone acetylation and expression from those genes. C – Transcriptional activation complexes from the distal enhancer and the promoter dimerise through Ldb1 dimerisation domain. GATA1 and TAL1 bind DNA, LMO2 stabilises this binding and recruits Ldb1. Images A & B from Kim & Kim (2013)[90], C from Love *et al.* (2014)[91].

Chromatin looping at the β-globin locus is mediated by interaction between two transcriptional activation complexes, one forming at the LCR and the other at the promoter of the expressed gene. These complexes include GATA1, TAL1, LMO2 and Ldb1. Along with GATA1, which is required to maintain the open chromatin state at the locus, TAL1 binding is necessary to recruit LMO2 and Ldb1, and looping is lost when TAL1 is knocked down[61,90]. Although KLF1 is not a component of this complex, it is required to stabilise binding at the LCR[57,92].

Ldb1 and LMO2 are non DNA binding components of the complex, LMO2 is required for stabilisation of the complex, and for recruitment of Ldb1, with the LIM domain in LMO2 binding to the LIM-Interacting Domain in Ldb1[58,93,94]. Ldb1 contains a self-association dimerisation

domain, and is the component of the complex responsible for mediating the interaction between the complexes at the two different genomic positions, directly binding to its counterpart in the other complex (Figure 1.4C)[95,96]. In Ldb1 knock out cells, chromatin looping can be rescued by fusion of the dimerisation domain to LMO2[95]. If fused to an artificial zinc finger DNA binding protein, the dimerisation domain is capable of reactivating the transcriptionally silenced γ-globin gene, by induction of forced chromatin interactions[97,98]. Ldb1 mediated chromatin looping is also involved in transcriptional activation of other erythroid activated genes, including KLF1, GATA1, TAL1 and LMO2[91,96,99].

1.3.3 DNA Methylation

DNA methylation is a common epigenetic mark, and when located at promoters is often associated with transcriptional silencing; recruiting chromatin modifying complexes, or interfering with transcription factor to DNA interactions.

While the β-globin locus has no bioinformatically predicted CpG islands, it has been found that in tissues expressing the globin genes, the promoters of the active genes have reduced CpG methylation, suggesting that DNA methylation is a relevant factor in the regulation of the locus[100]. This has also been shown in a transgenic mouse model, where DNA methylation was found to reduce expression of the foetal globin genes by 20 times[101]. Similarly, using a TALE-TET1 construct, targeted demethylation of four CpG sites in the β-globin promoter region is enough to reactivate the gene in K562 cells[102]. The maintenance DNA methyltransferase DNMT1 associates with BCL11A, and is required to maintain transcriptional silencing of the repressed β-globin like genes[103,104].

1.3.4 Histone Modifications

Histone modifications play a role in gene regulation, recruiting chromatin modifying complexes and signalling the transcriptional state of the gene. Histone marks are therefore very informative and useful to assay, especially in parallel with transcriptomic approaches such as RT-PCR or RNA-seq. They can be used to predict whether a gene is active or repressed, e.g. positive residues in the tail of histone H3 are commonly acetylated in euchromatin[9,105]. Methylation of lysine residues on histone tails indicates different states depending on which residue is methylated; H3K4 methylation is associated with active promoters, and H3K36 methylation is associated with actively transcribed regions, whereas H3K9, H3K27 and H4K20 methylation is

usually associated with inactive genes[9]. In the case of the β-globin locus it is important that the transitions between these marks by histone deactylases (HDACs) and acetyltransferases (HATs) as well as the methyltransferases and demethylases is controlled to regulate the complex expression pattern of the globin genes.

The active genes at the β-globin locus have high levels of H3K27 acetylation, and low levels of H3K27 methylation, and this is reversed in the transcriptionally silent genes[106]. At the β-globin promoter, GATA1 recruits the histone acetyltransferase CBP and NF-E2 recruits the histone methyltransferase MLL2, promoting histone acetylation and H3K4 methylation and activating transcription[54,107,108].

GATA1 and KLF1 are also required to maintain the chromatin organisation at the β-globin locus, and knockdown of KLF1 in K562 cells reduces expression of the γ-globin genes, as well as H3K9ac, H3K14ac & H3K27ac histone marks at both the LCR and γ-globin, and disrupts chromatin looping between the two sites[57]. Chromatin looping was also lost upon knockdown of CBP in these cells, demonstrating the importance of histone modifications in regulating the structural organisation of chromatin at the locus[90].

1.4 Erythroid Development

Haematopoiesis is a complex and tightly regulated process, by which a wide variety of cell types with highly specialised structures and functions are produced from a small pool of HSCs. Many haematopoietic cell types have a greatly reduced life span compared to other tissues, and this also varies greatly between the haematopoietic lineages. As a result, haematopoiesis is a continuous process, highly responsive to stimuli to allow coordination and maintenance of the many blood cell populations in the proportions required.

Erythropoiesis is the developmental pathway within haematopoiesis that results in the production of erythrocytes, and in adults is responsible for the generation of approximately $2x10^{11}$ red blood cells per day[109].

1.4.1 Normal Erythropoiesis

The major sites of haematopoiesis change throughout development. The initial haematopoietic populations reside in the embryonic yolk sac, and during foetal development haematopoiesis occurs in the liver and spleen. After birth, HSCs migrate to the bone marrow, which remains the major site for haematopoiesis throughout adulthood[109–111].

Since SCA does not affect the embryonic or foetal stages of development, when alternative β-globin paralogues are expressed, this section will focus on erythropoiesis in the bone marrow.

1.4.1.1 Haematopoietic Stem Cells & Early Stage Progenitors

A haematopoietic pool is maintained in the bone marrow, with self-renewal potential allowing expansion of the HSCs to replace those undergoing differentiation. Various cytokines in the bone marrow signal for quiescence of HSCs, as well as for their retention in the bone marrow through expression of adhesion molecules, including α4β1 integrin, which binds to VCAM1 expressed on stromal cells in the bone marrow[112–114]. CXCL12 is one of the key cytokines in this process, binding to CXCR4 expressed on HSCs, and is essential for maintaining quiesence[113,115,116]. Several cytokines such as Granulocyte colony stimulating factor, Flt3 Ligand and Interferon-α can be used to induce the activation and mobilisation of HSCs in mice[117,118].

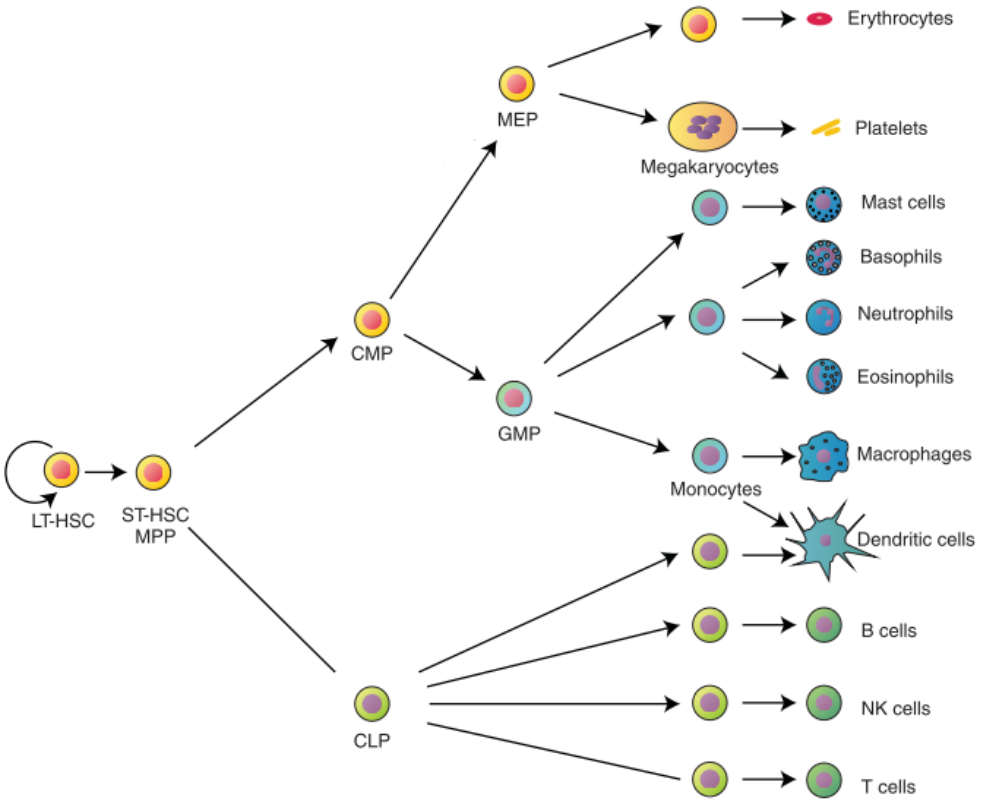


Figure 1.5: Simplified overview of haematopoietic development and terminally differentiated cell types produced from HSCs. MPP (Multipotent Progenitor), CLP (Common Lymphoid Progenitor), CMP (Common Myeloid Progenitor), MEP (Megakaryocyte-Erythroid Progenitor), GMP (Granulocyte-Macrophage Progenitor). Image adapted from Dzierzak & Philipsen (2013)[109].

The process of HSC development, and the potential cell types they can give rise to are summarised in Figure 1.5[109]. Upon haematopoietic stimulation, either as a result of natural blood homeostasis or in response to injury and blood loss, long-term HSCs divide asymmetrically, producing one long-term HSC and one short-term HSC, which no longer has capacity for self-renewal[119,120]. The short-term HSC develops into a Multipotent Progenitor (MPP), before committing as either a Common Myeloid Progenitor (CMP) or Common Lymphoid Progenitor (CLP)[121,122]. Before this stage the cells are capable of differentiating into any blood cell type, but CMPs are lineage restricted to myeloid cells, and CLPs can only produce the white blood cells[119,122].

CMPs develop into either Granulocyte-Macrophage Progenitors (GMPs), responsible for production of the granulocytes and macrophages, or Megakaryocyte-Erythroid Progenitors (MEPs). MEPs develop into either megakaryocytes, which produce platelets, or proerythroblasts, early stage erythroid progenitors that develop into the terminally differentiated erythrocytes[109,122].

1.4.1.2 Erythroblast Development

Erythroblast development occurs in the bone marrow at erythroblastic islands, where proerythroblasts cluster around central macrophages[109,123,124]. In humans this usually consists of 10-30 erythroid progenitors per macrophage[123]. Proerythroblasts develop through three distinct stages identifiable through cytology, before enucleation and final erythrocyte maturation, all of which occurs at the erythroblastic island.

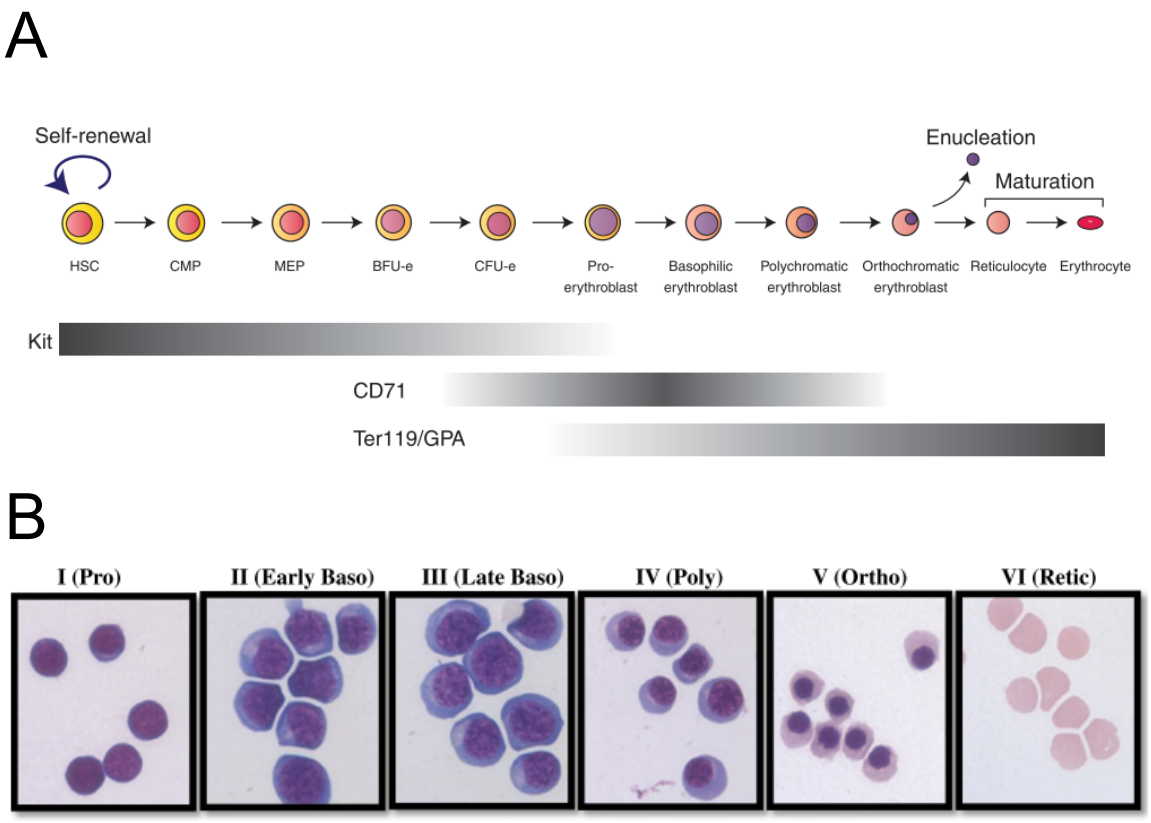


Figure 1.6: Summary of erythroid development, showing the individual erythroblastic stages, as well as the enucleation step. A – Cell surface expression is shown for Kit, CD71 and GPA (Ter119 is the murine equivalent of GPA). Kit expression is a marker for early stage progenitors, and is lost by the proerythroblast stage. CD71 is expressed during erythroblast development and is lost by the enucleation of the orthochromatic erythroblast. GPA is a late stage erythroid marker, with increasing expression levels during erythroblastic development, and is expressed highly on terminally differentiated cells. B – Cytology of erythroblasts isolated from human bone marrow. Image A is from Dzierzak & Philipsen (2013)[109]. B is from Hu *et al.* (2013)[125]

Proerythroblasts develop into basophilic erythroblasts, followed by polychromatic erythroblasts and finally orthochromatic erythroblasts before enucleation (Figure 1.6)[109]. Pro-erythroblasts have very little visible cytoplasm, and appear as small, tightly packaged cells. As they develop into basophilic erythroblasts the cytoplasm becomes visible, appearing to expand out from the nucleus. Basophilic staining in the cytoplasm is reduced as haemoglobin accumulates, and is accompanied by the appearance of large vesicles containing ferritin imported from the macrophage[126]. During the polychromatic erythroblast stage the nucleus condenses, and at the

orthochromatic erythroblast stage stains very dark. Rapid division occurs during this process, and the cell size decreases throughout these terminal stages[127,128].

It is interesting to note that during erythroid development each cell goes through the process of globin switching observed during embryonic and foetal development, transitioning from HbF to HbA[129–131].

1.4.1.3 Enucleation & Reticulocyte Maturation

The process of enucleation follows the orthochromatic erythroblast stage, and asymmetrically divides the cell into two products: a pyrenocyte, containing the condensed nucleus with a thin cytoplasmic shell, and a reticulocyte, consisting of the remaining organelles and the majority of the cytoplasm[132,133]. The pyrenocyte presents phosphatidylserine on the cell surface and is engulfed by the nearby macrophage at the erythroblastic island[134,135], while the reticulocyte matures, developing into an erythrocyte.

Enucleation is an important stage in erythropoiesis, and provides multiple structural advantages to the terminal erythrocyte, increasing flexibility and reducing both mass and volume, whilst maintaining its function as a haemoglobin rich oxygen transporter[136–138]. Enucleation also acts as an extremely effective developmental gate (there is no known mechanism by which a reticulocyte can re-absorb its nucleus), and given that erythrocytes are by far the most abundant cell type in the body, perhaps this irreversible block on cellular reprogramming is protective from an oncogenic perspective. It has been suggested that mitochondria are removed for a similar reason, since they have little influence over cell size or flexibility, but are a major source of oxidative stress[136].

1.4.2 Stress Erythropoiesis & Erythroid Progenitors in the Peripheral Blood

Under conditions of hypoxic stress, stress erythropoiesis is triggered. The process of stress erythropoiesis is not fully understood, but it has long been linked to increased HbF levels[139–142]. It is thought to be alternative erythropoietic developmental pathway, resulting in the rapid production of immature erythrocytes. During this fast-tracked differentiation process, the globin switch from HbF to HbA does not occur, resulting in an increased proportion of F-cells in circulation[143]. Due to the demonstrated ability of hypoxic stress to trigger F-cell production, and the cytotoxic effects of HU treatment, induction of stress erythropoiesis has been suggested as a mechanism of action for HU[143,144].

Stress erythropoiesis is triggered by a hypoxic response, mediated by the hypoxia induced transcription factor HIF2, which results in increased erythropoietin (EPO) production[145]. Induction of stress erythropoiesis is also dependent on functional EPO Receptor (EPOR) activation of STAT5[146]. Under stress erythropoietic conditions induced by phlebotomy in sheep, erythropoietin receptor (EPOR) levels are doubled, and mice haploinsufficient for EPOR are unable to elicit a stress erythropoietic response[147,148]. Stress response of dendritic cells has also been associated with activation of stress erythropoiesis, mediated through increased expression of Stem Cell Factor (SCF), the ligand for Kit[149].

SCA patients have increased levels of stress erythropoiesis as a result of hypoxia associated with the disease, and likely as a result of increased mobilisation, higher levels of stress erythroid progenitors have been observed in the peripheral blood[150].

1.4.3 *In vitro* Culturing of Erythroid Progenitors

Erythroid progenitors primarily reside in the bone marrow, but small populations have been identified in the peripheral blood of healthy individuals, and can be isolated and expanded in culture. There are many variations of these culturing techniques, and they are widely used, presenting a source of erythroblasts more easily accessible than by invasive bone marrow sampling.

In 1989 Fibach *et al.* demonstrated that it was possible to grow and differentiate erythroid progenitors isolated from the peripheral blood using a liquid *in vitro* culture[151]. Prior to this culturing was restricted to the establishment of macrophage dependent erythroblastic colonies on agar[152].

Human erythroid progenitors are commonly cultured from bone marrow, peripheral blood, cord blood and foetal livers, and in many protocols are enriched for $CD34^+$ cells prior to culturing, with the aim of purifying the early stage progenitor population[70,129,150,153–158]. However the benefits of this are disputed, with evidence suggesting that most of the erythroblastic growth potential actually comes from the $CD34^-$ population[159].

The timeline of the cultures varies greatly between different protocols, with some lasting up to 60 days[153]. This is mostly dependant on the design of the culture, with some single-phase cultures inducing differentiation from the start, two-phase cultures including an expansion phase before inducing differentiation, and three-phase cultures including two distinct expansion phases before differentiation[153,154,159].

The *in vitro* culture technique tested in this project was a two-phase liquid culture, performed on Peripheral Blood Mononuclear Cells (PBMCs) that were not $CD34^+$ enriched.

1.4.3.1 Culture Components

Various different media and serum are used in the *in vitro* culture systems, as well as a wide variety of different cytokines and growth hormones to stimulate proliferation and prevent apoptosis. All of these cultures include EPO, and the majority contain SCF, both of which have important proliferative and anti-apoptotic roles in erythropoiesis[148,160–162]. In addition, the culture tested in this project contains dexamethasone, Insulin-like Growth Factor 1 (IGF1) and Interleukin-3 (IL-3).

Dexamethasone is a corticosteroid that acts as a ligand for the glucocorticoid receptor. It has been shown to increase the longevity of erythroid progenitors in culture, promoting self-renewal and preventing terminal differentiation *in vitro*[161,163–165]. IGF1 is a growth hormone with a similar structure to insulin, and binds to the IGF1 receptor on the cell surface, signalling for survival and proliferation. Under culture conditions, IGF1 has been shown to increase erythroblast numbers, and is required for later stages of differentiation[166–168].

IL-3 is a cytokine that stimulates both growth and differentiation in haematopoietic populations, and in combination with EPO stimulates early stage haematopoietic progenitors to commit to the myeloid lineage. However, if IL-3 is present in the media for the duration of the culture, a large proportion of these progenitors are stimulated to terminally differentiate into mast cells, rather than erythrocytes[169]. Therefore culture medium is only supplemented with IL-3 during the first phase (the expansion phase), and not during the second phase (the differentiation phase).

1.4.3.2 Cell Surface Markers

Cell surface markers can be used to accurately identify the developmental stage of erythroid progenitors, either directly from the peripheral blood, or after *in vitro* expansion and differentiation.

CD71 and Glycophorin A (GPA) can be used to assess the stage of the erythroblasts. CD71 is a membrane receptor for transferrin and is expressed early on in development to allow iron accumulation prior to haemoglobin production. Expression is then lost towards the final stages of erythrocyte maturation[109,129,170](Figure 1.6). GPA is an erythroid specific glycoprotein that is

transcriptionally silent during early haematopoiesis, and is expressed at high levels after the basophilic erythroblast stage[109,129](Figure 1.6).

c-Kit is the receptor for the ligand SCF and is expressed at high levels on HSCs, where SCF binding promotes self-renewal and expansion of the HSC pool[171–173]. c-Kit is negatively regulated by GATA1, so is lost shortly after the switch from GATA2 to GATA1 early in erythroid development[173](Figure 1.6). CD34 is also an early HSC marker that is lost during haematopoietic development, before reaching the proerythroblast stage[174].

CD45 (aka leukocyte common antigen) is expressed from an early stage during haematopoietic development, but is lost from the erythroid lineage, and can be used as a leukocyte specific marker[125,175]. CD14 is a marker expressed highly on monocytes, and was used since monocytes are the largest source of contamination of the in vitro culture[175,176].

1.5 Sickle Cell Anaemia

1.5.1 History of Sickle Cell Anaemia

SCA was first recorded in the Western scientific literature by J. B. Herrick in 1910. The case was that of a patient from Grenada in the West Indies, with healthy parents and three healthy siblings, who presented "unusual blood findings" and red blood cells with a "large number of thin, elongated, sickle-shaped and crescent-shaped forms"[177]. Two additional cases were published in 1911 and 1915, and in 1922 a fourth patient was presented by V R Mason, with the first reference to the disease as "Sickle Cell Anaemia"[178–180].

The heritability of the disease was confirmed by two separate publications in 1949, where sickle trait (also referred to as sicklemia or drepanocytosis) was suggested to present in a dominant pattern in heterozygous carriers of the disease, while only homozygous cases had the full clinical presentation of SCA[181,182].

In 1949 Linus Pauling *et al.* noted a shift in the electrophoretic mobility of haemoglobin from SCA patients, correctly hypothesising that the sickle haemoglobin has a net positive charge of 2-4 ions more than healthy haemoglobin[183]. He went on to hypothesise that these haemoglobin molecules "might be capable of interacting with one another at these sites sufficiently to cause at least a partial alignment of the molecules within the cell", which remains the widely accepted hypothesis of SCA disease aetiology, more than 65 years after publication[183]. Pauling's 1949 paper is considered a landmark publication in the field of molecular medicine, being the first case of a disease being attributed to a specific molecular defect, and SCA became the first ever "Molecular Disease"[183].

Pauling's discovery of the differing charge on the sickle haemoglobin molecules was confirmed in 1956, with the discovery of a substitution of a negatively charged glutamic acid residue for a neutral valine residue on each of the two β-globin subunits[184].

In 1978, a genetic association was identified between the sickle variant and a disrupted restriction digest site, 5kb downstream from the β-globin gene[185]. This formed the basis of using restriction fragment length polymorphisms (RFLPs) as a diagnostic tool, and led to antenatal screening for SCA in affected families[186,187].

1.5.2 Sickle Cell Disease Epidemiology and Malarial Resistance

Due to the severity of SCD and the fact that it is a genetic disorder, it might have been expected that the disease would be less prevalent as a result of negative selection, since infant mortality is high without early intervention. The prevalence of the sickle genotype is increased because individuals who are heterozygous for the HbS allele have increased resistance to severe malaria from Plasmodium falciparum[188,189]. It is therefore likely that the prevalence of the sickle cell genotype is maintained in part by a balance of the positive selection for heterozygosity against the negative selection for homozygosity. This also explains why the disease is most prevalent in tropical and sub-tropical regions where malaria is endemic, including sub-Saharan Africa, the Mediterranean, the Middle East and India (Figure 1.7)[190]. The global frequency of the HbS allele as estimated by the 1000 Genomes Project is 0.03, this is increased to 0.10 in African populations, and 0.14 in the Yoruban subpopulation in Nigeria[191].

Individuals who are heterozygous for the HbS allele do not present with the SCA phenotype, but do have an increased protection from the symptoms of malaria. However, patients homozygous for the sickle HbS allele do not have this protective effect, they experience the severe anaemia and autosplenectomy associated with SCA, and are actually predisposed to an increased risk of death from malarial infection[192,193].

Carriers of the HbS allele are not actually protected from the malarial infection, but from the symptoms. The proportion of erythrocytes infected with P. falciparum appears to be reduced in these individuals, believed to be at least partially due to increased rates of sickling in infected erythrocytes, and the subsequent clearance of these cells from the bloodstream by phagocytosis[193–195]. It has also been shown that cytoskeletal disruption in erythrocytes of HbS carriers prevents correct trafficking of the malarial proteins to the cell membrane, which are required for endothelial adhesion and thought to increase parasite survival[29,193,196].

The sickle mutation is believed to have arisen independently multiple times, in different populations under strong selective pressure by malarial infection. This is demonstrated by the different genetic haplotypes that associate with the β^S mutation, as shown in Figure 1.8[187].

Malaria is not endemic to Northern Europe, and so the selective advantages of the sickle mutation are minimised in these countries. However, likely as a result of increased global migration over the last century, SCD is the most common severe genetic disorder in the UK, and affects approximately 13,000 individuals[197–199]. 140-175 babies are born with SCD in England each year, equating to 0.22-0.28 per 1000 births, this is increased to 5.6 and 14.7 in

patients with Caribbean and African ancestry respectively[45]. A published analysis of SCD treatments and sickle related hospitalisations in 2010-2011 estimated that SCD costs the NHS roughly £18.8 million annually, with over 6,000 hospitalisations due to acute sickle pain crisis alone[198].

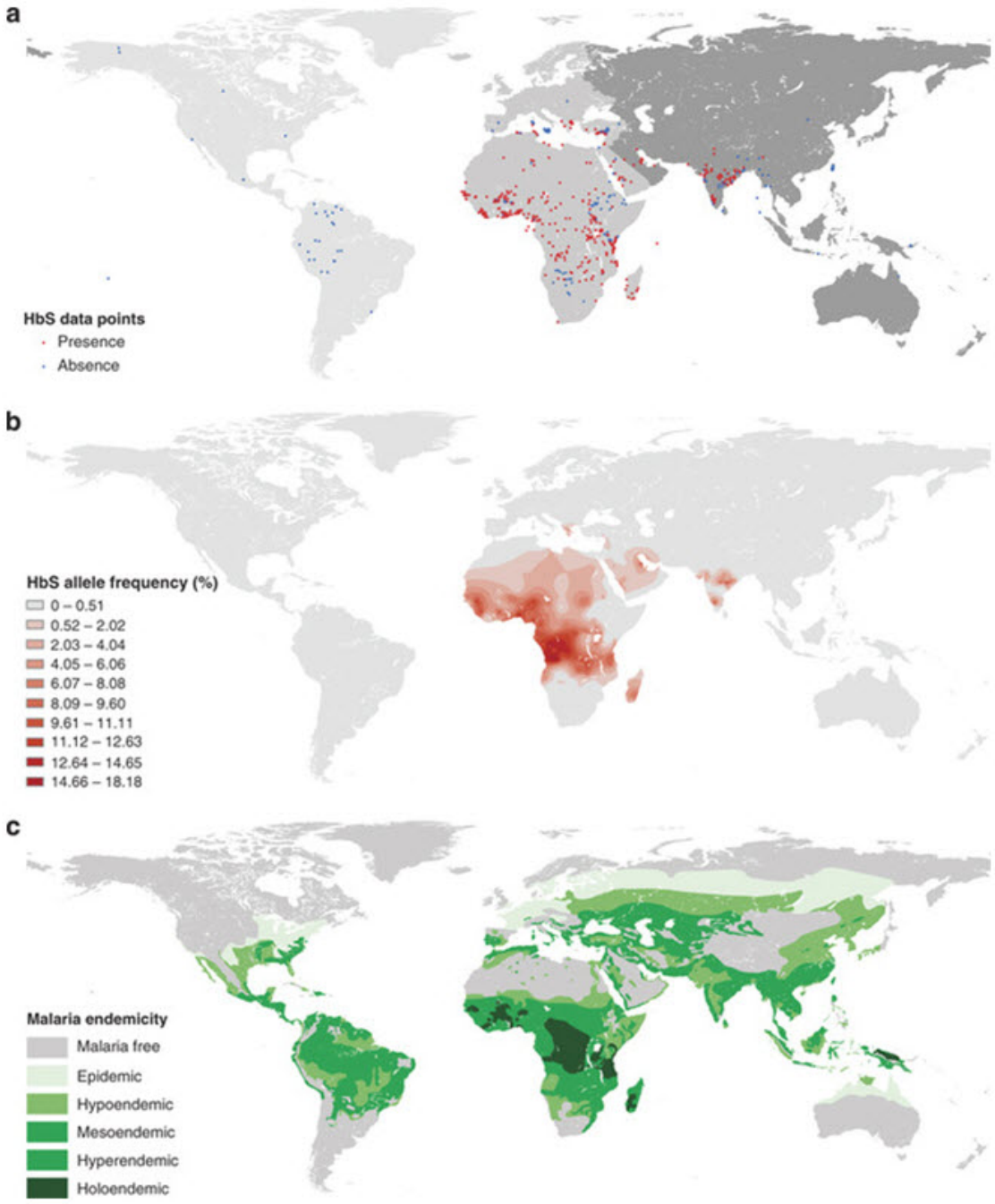


Figure 1.7: Diagram illustrating the global distribution of (A & B) HbS allele, (C) Plasmodium falciparum infections. Note the strong overlap in central Africa. Image from Piel et al. (2010)[190]

Figure 1.8: Map showing the distribution of different haplotypes that associate with sickle globin alleles. The sickle globin mutation is believed to have arisen independently multiple times across malaria affected countries in Africa and Southern Asia. Image from Gabriel & Przybylski (2010)[187].

1.5.3 Sickle Cell Anaemia Symptoms

The major symptoms of SCA arise from severe chronic haemolytic anaemia and acute vaso-occlusive events. Hypoxia due to anaemia causes cell death in organs and peripheral tissues, while recurrent episodes of vaso-occlusion prevent blood flow to large areas of tissue, causing increased cell death, and resulting in many symptoms, including chronic and acute pain, stroke, chronic lung disease, osteonecrosis, renal failure and retinopathies[25,200].

Sickle cell crisis is an inflammatory response triggered by vaso-occlusion, and is the most common reason for hospitalisation of SCA patients[198]. Evidence suggests that damage caused to the endothelium by rigid cells triggers this response, and activated endothelial cells, as well as neutrophil extracellular traps (NETs) are involved in this process[201–203].

Autosplenectomy is a common occurrence in SCA patients, with splenic function often severely impaired from a young age (<12 months), due to damage caused by vaso-occlusive events and continuous filtering of abnormal sickled cells from the blood[204,205]. Loss of splenic function leaves SCA patients susceptible to infections[206].

Current risk categories for stroke in SCA children are assessed by measuring blood flow in the brain using transcranial Doppler ultrasonography (TCD), and patients with a flow of <170cm/s are considered at low risk, however a study found a significantly high incidence of intracranial stenosis as well as silent cerebral infarcts in these patients, both of which are predictive markers of future stroke. This shows that the current methods used to predict severity of SCA symptoms can be improved[207].

Despite the well-characterised genotype and the classification as a classical Mendelian recessive disorder, SCA pathology varies greatly between patients, both in terms of severity and symptoms observed. Patients with the same β-globin mutation will in some cases have high incidence of stroke and clotting from an early age, and others will be largely unaffected, reaching old age without incident; this implies that there is an additional component to the disease[200,208].

Symptoms observed in a longitudinal study of 1056 patients

Acute Clinical Events	
Symptom	*% Patients*
Hospitalisation:	76
Sickle Related	73
Painful Sickle Crisis	70
Associated Sickle Crisis	51
Acute Chest Syndrome	48
Hypersplenism	20
Bone Infarction	15
Aplastic Crisis	14
Trauma	13
Meningitis/Septicaemia	12
Neurologic Disorder	12
Dactylitis <4 years	4

Irreversible Organ Damage	
Symptom	*% Patients*
Gall Bladder Disease	28
Avascular Necrosis	21
Sickle Chronic Lung Disease	16
Leg Ulcer	14
Priapism (<17 yrs)	14
Renal Failure	12
Cerebrovascular Accident	11
Retinopathy	9
Death	*232 Total*
Sickle Caused	170 (73%)
<20 years	46 (20%)

Table 1.1: Table showing data from a longitudinal study of 1056 SCA patients over 40 years, adapted from Powars et al. (2005)[200]. A large variety of symptoms are presented, and the percentage of patients that presented with each clinical event is shown, the study found that patients that present a chronic clinical event are more likely to have future events as well.

A longitudinal study of 1056 patients over 40 years is summarised in Table 1.1. The table shows the variety of symptoms, and the study found that many of the chronic clinical events were risk factors for future clinical events, further demonstrating that the severity of symptoms varies from patient to patient, despite sharing the same sickle-globin mutation[200]. This wide variety of symptoms provides a challenge in clearly defining boundaries between severe and mild patients

for our analyses, since phenotypic severity is a sliding scale. For this study, patients that suffer a stroke or serious clinical incident before the age of 18 are considered as severe, and patients that reach 30 years without serious incident as mild. In addition to this, onset of symptoms usually associated with old age (e.g. retinopathies & hypertension) at <18 years, is considered severe.

1.5.4 Treatments

Due to the variation in phenotype severity observed, there is no 'one size fits all' treatment for SCA. A variety of treatment options are available, with different advantages and disadvantages, and a specific treatment plan is generally decided on a case by case basis depending on the frequency and severity of the symptoms observed.

Bone marrow transplant remains the most comprehensive treatment option, at least partially replacing the host's sickle HSC population with that from a healthy donor, leading to increased production of healthy erythrocytes[209–211]. However, bone marrow transplantation is an invasive procedure, requiring the availability of an HLA-matched sibling, and risks severe complications such as graft versus host disease[212–214]. As a result, very few SCA patients receive this treatment.

Blood transfusions are an effective temporary treatment option for SCA patients, increasing the proportion of healthy erythrocytes circulating in the peripheral blood. These can either be administered as treatment in response to an acute clinical event, to alleviate symptoms and assist in recovery, or they can be used as a preventative measure for patients at risk of stroke, as identified by severe clinical history or TCD[215–218]. The STOP (Stroke Prevention in Sickle Cell Anaemia) trial demonstrated that regular transfusions as a long-term treatment plan was able to reduce incidence of primary stroke in children with SCA[218].

Patients receiving regular blood transfusions require frequent hospital visits, and risk long term complications. Regular blood infusions increases the risk of alloimmunisation, triggering an adaptive immune response to donor erythrocytes and resulting in haemolysis of transfused blood, reducing the effectiveness of treatment and increasing the burden of clearing already high levels of haemolytic debris from the bloodstream[219]. Interestingly the rates of alloimmunisation vary between countries, and is thought to be linked to the shared genetic ancestry of the donor and the recipient, i.e. patients of African or Asian ancestry treated in countries with a majority of white donors experience higher rates of alloimmunisation[219–221].

As well as increased risk of blood borne infections, regular blood transfusions increase levels of iron, which is released from erythrocytes upon haemolysis, and over time can result in iron overload[218]. Due to the redox activity of unbound iron, iron overload results in increased free radical production and higher levels of oxidative stress and inflammatory response to the damage caused[222]. Excess iron that cannot be excreted builds up in hepatocytes and eventually cardiomyocytes, and can lead to cirrhosis and heart failure[222,223]. Iron overload as a result of regular transfusions is a significant cause of death among SCA patients, and the monitoring of iron levels is important to minimise risks, and to ensure that the need for chelation therapy is identified at an early stage[222,224,225].

Hydroxyurea (HU, also referred to as hydroxycarbamide) is the most commonly prescribed drug for treatment of SCA, and significantly reduces frequency of sickle cell crises and hospitalisations, increases overall survival and is preventative of strokes and other vaso-occlusive events[226–230]. HU has been used as a chemotherapy drug to treat myeloproliferative neoplasms (MPNs) since the early 1960s, and was first tested for use in SCA patients in 1984, having being shown to increase foetal haemoglobin levels in anaemic monkeys[231–233]. MPNs are myeloid lineage cancers that result in uncontrolled expansion of blood cell populations, and HU was used for treatment since it disrupts DNA replication, and inhibits proliferation of the malignant cell types[231,234,235]. It is now known that HU blocks DNA synthesis through inhibition of Ribonucleotide Reductase (RNR), which catalyses the reduction of ribonucleotides to the deoxyribonucleotides required for DNA strand elongation, either during replication or DNA repair[236,237].

In SCA patients this cytotoxic effect of HU treatment results in a reduction of reticulocyte and white blood cell counts[22]. However, rates of haemolysis are decreased and a greater proportion of these reticulocytes survive, and the reduction in white blood cells may be beneficial in improving blood flow through narrow blood vessels[238].

The main benefit of HU treatment in SCA patients is mediated through the increase in foetal haemoglobin (HbF) levels; increased HbF provides a functional alternative to the sickle globin, as well as diluting the intracellular HbS concentration, competing for binding and forming hybrid tetramers ($\alpha_2\gamma\beta^S$). This is mediated through repression of the MYB, BCL11A, KLF1 and TAL1 regulatory pathways, although how this is achieved is not fully understood[81,239,240]. Evidence supports the involvement of miRNA as well as the nitric oxide (NO) and cGMP signalling pathway, with HU treatment known to increase NO production either through signalling or as a

by-product of its degradation[239,241–245]. This increase is also caused by an increase in the proportion of F cells (erythrocytes with high HbF) in the peripheral blood, which occurs as a result of increased stress erythropoiesis in response to the cytotoxic effects of HU (1.4.2)[246,247].

Various other processes are also thought to be involved in the mechanism of action of HU, including increasing the oxygen affinity of haemoglobin (only the deoxygenated state of haemoglobin polymerises), mediated through a reduction in adenosine signalling[248]. Another proposed mechanism is through altering the expression of cell surface adhesion molecules, reducing rates of vaso-occlusion[249].

Because of the many processes and pathways that HU appears to affect, it is difficult to identify distinct mechanisms of action, since it is not clear which of these changes occur as a direct result of HU interaction, and which occur as a secondary response to either the cytotoxic stress of HU treatment, or to the general improvement in disease phenotype. For this reason it is important that reliable protocols are in place to investigate the epigenetic and transcriptional changes that occur in specific homogenous erythroid populations from SCA patients *in vivo*, since the additional stress of *in vitro* culturing further masks the pathways being influenced by HU treatment. While most current studies either use mouse models or *in vitro* culturing techniques (discussed in 1.4.3), a study by Walker *et al.* used a magnetic bead separation technique to isolate both DNA and RNA from erythroid progenitors in patients undergoing HU therapy[245].

Trials comparing HU treatment and phlebotomy to chronic blood transfusions and chelation therapy have been conducted in the USA, looking to prevent complications such as iron overload as a result of regular transfusions. The SWiTCH (Stroke With Transfusions Changing to Hydroxyurea) study was carried out in SCA patients that had experienced stroke at a young age (<18), and were receiving regular transfusions to prevent a secondary stroke[250,251]. TWiTCH (TCD With Transfusions Changing to Hydroxyurea) was a similar study, investigating SCA patients receiving regular transfusions to prevent initial stroke, having been identified as at risk by TCD[252]. SWiTCH was terminated early when HU was shown to be inferior to transfusions in preventing adverse effects, and did not alter liver iron levels[250]. TWiTCH however did demonstrate that liver iron levels were reduced in the HU treatment arm, and showed non-inferiority to transfusions in terms of effect on TCD velocity and risk for stroke[252].

Prior to HU, other drugs had been used to re-activate expression of foetal haemoglobin, including butyrate and 5-azacytidine[253–256]. 5-azacytidine is incorporated into the genome as a

homologue of cytosine that cannot be methylated, inhibiting DNA methylation. This was thought to reactivate γ-globin by removing repressive DNA methylation marks at the promoter, highlighting a role for epigenetic factors in the pathology of SCA[254,256]. The study by Walker *et al*. found that HbF induction in response to HU was not accompanied by hypomethylation at the γ-globin promoter[245]. Interestingly butyrate is thought to increase translation efficiency of γ-globin mRNA, suggesting the role of post-transcriptional regulators[257]. More recently, novel HbF inducers have been proposed, including rapamycin and pomalidomide[258,259].

Anti-sickling agents such as 5-hydroxymethyfurfural (5-HMF) has also been suggested as a potential treatment option, increasing the oxygen affinity of haemoglobin, and therefore reducing sickling[260,261].

The vasodilatory effect of NO signalling has been identified as a potential mechanism by which HU may ameliorate SCA symptoms severity, and arginine treatment has been shown to reduce pain during sickle cell crises through the same mechanism[242,262,263]. NO is a vasodilatory signalling molecule that is synthesised from L-arginine by NO synthases, and supplementing arginine levels was suggested as a way to increase NO levels[262,264]. However the importance of the role of NO in SCA is disputed[23], and the study that found arginine therapy to relieve pain during sickle crises also observed that low arginine in these patients didn't always correlate with low NO, and suggested that the benefits of arginine therapy may also be conferred through NO independent pathways[262].

As a result of autosplenectomy, SCA patients have an increased susceptibility to infections, and preventative measures are started at an early age, with patients receiving additional vaccinations for Streptococcus pneumoniae, Influenza and Hepatitis B, as well as prophylactic antibiotic treatment in children[265–268].

1.6 Known Genetic Modifiers of SCA Phenotype Severity

Haplotype analysis of the β-globin locus in SCA patients in Africa and Southern Asia has revealed a correlation between haplotype and severity of symptoms, highlighting that genetic factors outside of the β^S mutation play a role in disease pathology[187,269].

Genetic variants in genes and pathways influencing the pathophysiology of SCA can account for some of the variation in severity that is observed. The two best characterised mechanisms of action for these genetic modifiers are those that increase HbF production, and those that cause α-thalassaemia. Both of these play a role in ameliorating the severity of the SCA phenotype.

1.6.1 Foetal Haemoglobin

Foetal haemoglobin (HbF) is expressed during foetal development, with a switch to HbA occurring shortly after birth. HbF is made up of two α subunits and two γ subunits ($\alpha_2\gamma_2$), with β-globin transcriptionally silent at this stage. This explains why there are no *in utero* complications of SCA (other than those caused by SCA in the mother[270,271]), and why symptoms are not observed until several months after birth[272]. Even moderately increased γ-globin levels in an erythrocyte leads to competition for α-globin binding, reducing HbS formation, and polymerisation.

Errors can occur in the regulatory switch from γ-globin to β-globin, and in healthy adults individual erythrocytes may contain high HbF, these are referred to as F cells, and generally account for <1% of total blood haemoglobin[273,274]. Genetic variants have been identified that result in much higher levels of HbF in adults, collectively these are characterised as Hereditary Persistence of Foetal Haemoglobin (HPFH) syndromes. In HPFH patients, HbF expression is not limited to F-cells, but is expressed universally across erythrocytes, since rather than being a stochastic phenomenon caused by dysregulation on a cell by cell basis, every cell contains the same genetic variant, which will affect HbF γ-globin expression in the same way[269].

HPFH can be caused by genetic variants at the β-globin locus that directly influence gene expression. These include variants that disrupt the β-globin promoter, variants in the promoter or intronic regions of the γ-globin genes that are associated with transcriptional repression, and disruptions in the LCR, which interacts with promoters in the locus through long-range chromatin looping[274–276].

Alternatively genetic variants causative of HPFH can disrupt genes in pathways that regulate expression from the β-globin locus, these include genes such as BCL11A and KLF1[63,74,276,277]. Another mechanism for HPFH is through variants that result in increased haematopoiesis[276]. Under hypoxic conditions stress erythropoiesis can occur, stimulating an increased rate of erythrocyte production, achieved by the release of premature erythroid progenitor cells from the bone marrow[141,142]. Since the γ-globin to β-globin switch occurs during erythroid maturation, these premature erythrocytes still synthesise γ-globin, leading to an overall increase in blood HbF levels[139,143,278] (discussed in more detail in 1.4.2). MYB is haematopoietic regulator that is required for cell cycle progression of erythroid progenitors and HSCs, playing a crucial role in proliferation and differentiation[78]. As a result of this MYB has oncogenic potential, and variants up-regulating the function of MYB are implicated in various cancers, including leukaemia[279,280]. MYB acts as a transcriptional activator of key erythroid transcription factors KLF1 & LMO2, which both play a role in transcriptional regulation at the β-globin locus[77]. Variants in the intergenic region between MYB and HBS1L, have been linked to high HbF levels by GWAS, and ChIP-seq combined with Chromosome Conformation Capture (3C) shows a distal enhancer interaction with the MYB promoter[79,80,281].

Elevated HbF levels in healthy individuals is mostly asymptomatic. However, in patients with SCA or any other β-globinopathy, coinheritance of HPFH provides a functional alternative to β^S and results in a less severe disease phenotype.

HbF levels in SCA patients vary greatly, and associate with the different sickle globin haplotypes (Figure 1.8). Individuals with the Senegal or Saudi Arabia/India haplotypes generally have higher levels of HbF, and present a milder disease phenotype than those with the Bantu or Cameroon haplotype, while patients with the Central African Republic haplotype generally have the lowest HbF levels and present the most severe symptoms[269,282].

Additional genetic studies have identified three quantitative trait loci (XmnI-Gγ on chromosome 11p, BCL11A on 2p and the HBS1L-MYB intergenic region on 6q) that have been attributed to up to 50% of common HbF variation in SCA patients[74,283].

1.6.2 α-Thalassaemia

A-Thalassaemia is defined by insufficient production of the α-globin component of adult haemoglobin, resulting in a haemolytic anaemia. Similarly to γ-globin, α-globin is encoded by two paralogous genes (HBA1 & HBA2) both of which are situated in the α-globin like gene locus

(referred to as α-globin locus), shown in Figure 1.1. As is observed at the β-globin locus, developmental stage specific regulation results in a switch from the embryonic ζ-globin to adult α-globin, further adding to the repertoire of haemoglobin variants encountered during healthy development. Interestingly, it has been shown that $\zeta_2\beta^s_2$ haemoglobin does not polymerise in vitro, and expression of ζ-globin in a SCD mouse model reversed the phenotype, suggesting that similarly to γ-globin, persistent expression of ζ-globin could have a therapeutic effect on the SCA phenotypes.

The expression of four functional α-globin genes in adults makes α-thalassaemia a particularly variable disease, both in terms of genotype and phenotype. Expression levels from each individual gene vary with genetic polymorphisms in their regulatory regions, while frameshift and stop gain mutations prevent the translation of functional products, however the most common genetic cause is the large scale deletion of an entire α-globin paralogue[284,285]. Disease severity varies with the amount of functional α-globin, which depends on a combination of gene copy number and functional expression level from each gene present[285,286].

Loss of a single α-globin gene (αα/α-) is asymptomatic and is classified as being a silent carrier, and loss of two copies (α-/α- or αα/--) is referred to as α-thalassaemia trait, only presenting a mild phenotype (mild anaemia). In the presence of less than two functional α-globin genes, the symptoms become much more severe, resulting in symptoms including severe haemolytic anaemia, oedema, jaundice and skeletal/cardiovascular malformations[285,286].

Similarly to SCD, α-thalassaemia has been associated with protection against malaria, and in α-thalassaemia patients this effect is observed in both the heterozygous and homozygous forms[287–289]. This is contrary to what is observed in SCD, where the protective effect is limited to the heterozygous patients[193].

As a result of both disease phenotypes being evolutionarily advantageous in malaria rich regions, the global distributions of both SCA anaemia and α-thalassaemia are very similar, and they are often co-inherited, with 30-35% of SCA patients from an African ethnic background also having α-thalassaemia[290–293]. When co-inherited with even the silent carrier state of α-thalassaemia, HbSS patients have reduced sickling, resulting in reduced haemolysis and increased erythrocyte lifespan. As such, some of the major symptoms of SCA are ameliorated, including risk of stroke and acute chest syndrome, however reduced haemolysis in these patients also leads to increased blood viscosity and been associated with increased vaso-

occlusive crises and osteonecrosis[294–300]. These patients also have a reduced response to treatment with HU[301].

This effect on disease severity is most likely caused by a reduction in formation of the haemoglobin tetramer ($\alpha_2\beta^S_2$) due to a lack of α-globin. This results in lower intracellular HbS levels, and since β^S-globin is only pathogenic when incorporated into haemoglobin, there is a reduction in the likelihood of polymerisation reaching the threshold required to distort the erythrocyte membrane.

1.6.3 Epigenetic Modifiers

As well as variation in severity between individuals of different genotypes, it has been observed that the SCA phenotype can vary between monozygotic twins[302–304]. Since these individuals are genetically identical, it can be assumed that any variation is due to epigenetic or environmental factors, with phenotypic discordancy caused by the regulation of modifier genes rather than sequence variants.

1.7 Genetic Editing: CRISPR-Cas9

Recent advances in genomic editing techniques now present exciting opportunities for researchers investigating genetic disorders, not only for possible applications in the laboratory, but also for the potential to correct disease causing mutations in patients, which would become the pinnacle of personalised medicine. Unfortunately however, in addition to the ethical issues that will almost certainly be raised, there are technical obstacles that need to be overcome before this can become a reality.

Firstly, from a practical viewpoint, an effective delivery system is needed, capable of introducing the CRISPR machinery into all cells of a target tissue. This is more achievable for tissues with a high turnover rate such as the haematopoietic lineages, where the desired mutation only has to be introduced into the HSC pool in bone marrow. However, in developed tissues with low turnover rates a much larger number of cells will need to be targeted directly. For the majority of cases, genomic correction may need to be carried out at early developmental stages to be effective, either to correct a progenitor population before the tissue becomes established, or because correcting the genotype may not reverse any damage already caused by deleterious mutations in developed tissues.

Also, as is clearly demonstrated by the results in this study, the technique is not yet 100% accurate, especially when it relies on the use of the host's endogenous DNA repair machinery. While directed cleavage of DNA by CRISPR is efficient, the genomic integration of template sequences is less so, often resulting in introduction of small insertion or deletions around the target site. Off target effects as well as these aberrant on target effects of CRISPR therapy could be catastrophic and unpredictable, and unlike the majority of non-surgical medical treatments, irreversible. Ideally cells need to be isolated, edited, rigorously tested and replaced.

Bone marrow is one of the few tissues that can realistically fill all these criteria: blood cells have a high turnover rate, and therefore editing could potentially be done at any stage. Cells are repopulated from an easily isolatable stem cell population. Bone marrow transplants are already a commonly performed technique in severe haemoglobinopathies, and the dangers of host rejection are significantly reduced since the recipient's own marrow can be used.

1.7.1 CRISPR-Cas9 Discovery

The name 'CRISPR' (**C**lustered **R**egularly **I**nter**S**paced **P**alindromic **R**epeats), has no relevance to either the structure or mechanism of the CRISPR-Cas9 systems used in laboratories today, but instead refers to the structure of a highly repetitive locus that was initially identified in *E. coli* by Ishino in 1987, and was later observed in many other prokaryotes, often occurring at multiple sites in the genome[305–307]. The term 'CRISPR' was first coined by Jansen in 2002, along with the Cas (**C**RISPR-**as**sociated) family of genes that co-localised with the CRISPR loci[307].

In early 2005, both Mojica and Pourcel independently published that the non-repetitive regions separating these repeats (protospacer regions) were homologous to DNA sequences found in bacteriophages and other transmissible genetic elements, and they hypothesised that CRISPR loci play a role in protecting prokaryotes from pathogenic foreign DNA[308,309].

Also in 2005, Bolotin identified a variation in *Streptococcus thermophilus* and *Streptococcus vestibularis*, where Cas1-4 from previously investigated bacterial strains were absent, but had been replaced by a Cas1 homologue and two additional genes, one of these genes was Cas9 (although referred to in this paper as Cas5)[310]. This was the first discovery of what would later become known as the Type II CRISPR System, the system currently used in laboratories around the world[310,311].

In 2007 Barrangou *et al.* confirmed the theory that CRISPR functions as a bacterial immune system, demonstrating in *S. thermophilus* that when cultured in the presence of bacteriophages, some cultures acquired resistance, and that this correlated with the incorporation of bacteriophage genomic DNA sequence into new protospacer elements in the bacterial genome[312]. Additionally, creating knockout strains of Cas7 and Cas9 (referred to as Cas5 in the paper) they showed that Cas7 was required for incorporation of new protospacer elements, and therefore acquisition of resistance, and that Cas9 was required for the resistance to be effected[312]. The papers by Bolotin & Barrangou also both mention that Cas9 contains nuclease activity, and that it is likely important for its function[310,312].

In 2008, van der Oost *et al.* found that the entire CRISPR loci of repeating elements was transcribed as one long CRISPR RNA (crRNA), and that this was cleaved by a complex of Cas proteins into short guide RNA (sgRNA), each containing one of the protospacer sequences[313]. They cloned components from a Type I CRISPR system from *E. coli* into a CRISPR negative *E. coli* strain, and designed an artificial crRNA to target λ phage. This was the first example of a CRISPR system being artificially targeted for a specific sequence of interest[313]. Using this

system Oost *et al.* demonstrated that the CRISPR system was driven by sgRNA, cleaved from the longer crRNA transcript, and that this works for both sense and antisense RNA sequences, suggesting that dsDNA is the target. This theory was confirmed later on 2008 by Marraffini & Sontheimer, targeting an untranslated region of a conjugation plasmid, in the summary of this paper they go on to postulate that if the functionality of CRISPR is not limited to its use in bacterial systems, then it would have many potential uses, including its use in the clinic[314].

In 2010 it was shown that the cleavage site is not random, and that a double strand break is introduced 3bp upstream of the Protospacer Adjacent Motif (PAM), a sequence in the target DNA immediately at the 3' of the protospacer sequence[315].

In 2011, performing RNA-seq on *Streptococcus pyogenes*, Deltcheva *et al.* came across a small non-coding RNA adjacent to the CRISPR locus[316]. This tracrRNA (trans-activating CRISPR RNA) was found to be required for crRNA processing into sgRNA, and has a 24bp sequence complementary to the CRISPR repeating region, suggesting a mechanism where the crRNA cleavage is directed by binding of tracrRNA[316]. It was shown in 2012 that this tracrRNA also interacts with Cas9 through its secondary structure, and therefore acts as the link that physically connects the sequence targeting sgRNA to the nuclease activity of Cas9[311,317].

In 2012 two separate papers were published demonstrating significant optimisation in the use of the CRISPR-Cas9 System, Gasiunas *et al.* demonstrated that by purifying the individual components they were able to perform CRISPR mediated cleavage *in vitro*, directed by 20bp protospacer sequences in the crRNA[318]. Jinek *et al.* (published as a collaboration between the Doudna and Charpentier laboratories) demonstrated that the tracrRNA was required to interact with Cas9, and that it was possible to fuse the sgRNA and the tracrRNA into a single gRNA molecule, streamlining the process for use outside of bacterial systems[319].

In January 2013 the first examples of CRISPR use in human cells were published, initially by Zhang and Church separately (in the same issue of Science)[320,321], and then a few weeks later separately by Doudna and Kim[322,323]. All four of these papers used systems based on the *S. pyogenes* Type II CRISPR System, and the authors of all four papers have separately applied for patents covering various aspects of their work.

1.7.2 Mechanism: Bacterial 'Immune System'

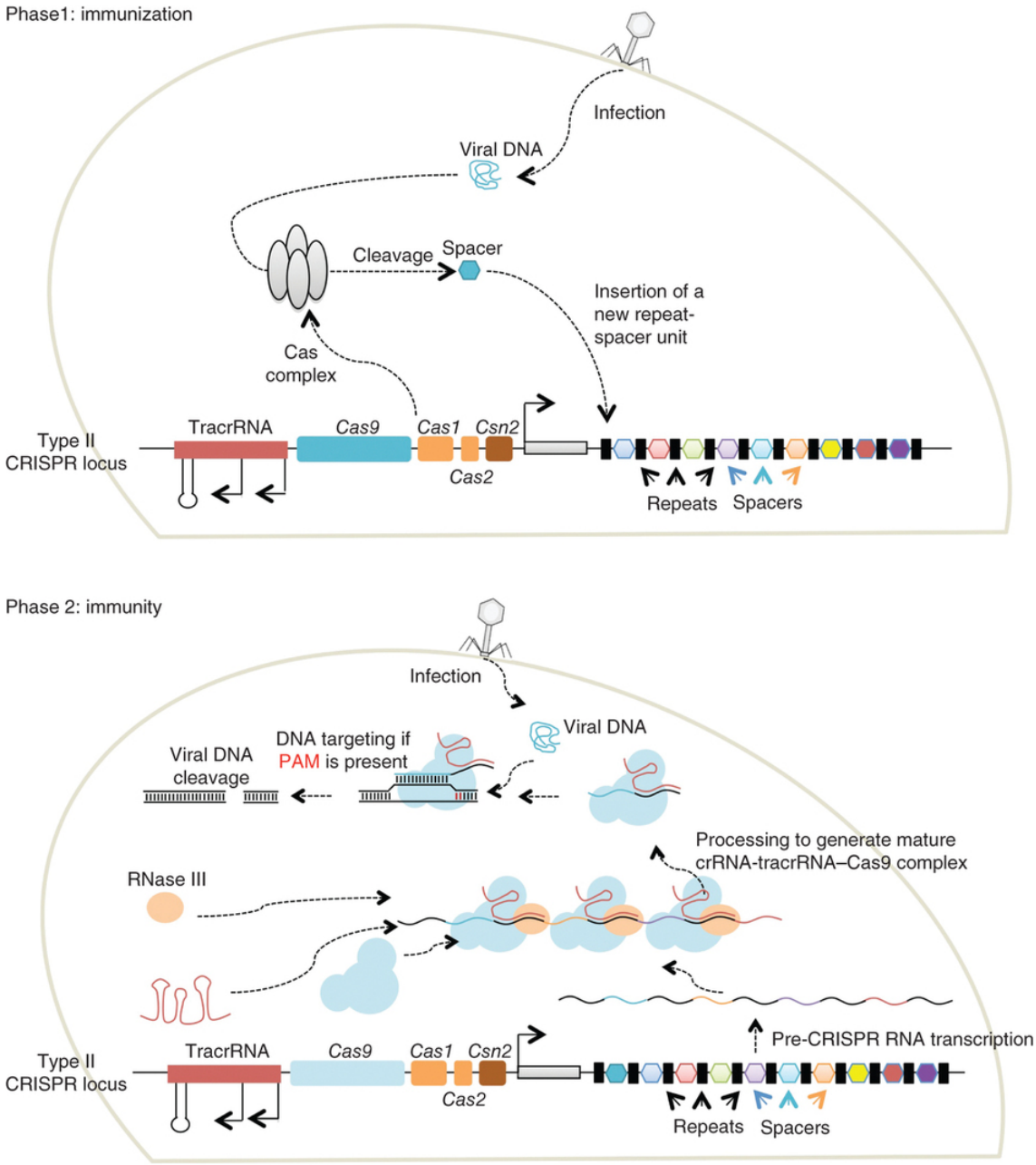


Figure 1.9: Figure from Mali *et al. (2013)*[324]. CRISPR Cas9 Type II System, showing the two distinct phases of bacterial 'immune response' and acquisition of resistance against invading viral DNA. Phase 1: Cas proteins (and Csn2) bind and recognise foreign DNA and cleave it into short 30bp 'spacers', and integrates these spacers into the host genome, at the 5' end of the CRISPR array, separated by 36bp repeats. Phase 2: the CRISPR array is transcribed in full, and tracrRNA recognises and binds to the repeat regions, directing RNase III cleavage of the crRNA into sgRNA. tracrRNA-sgRNA complex recognise and bind to homologous sequence on foreign DNA. Cas9 is recruited by tracrRNA secondary structure, and cleaves the target DNA.

The *S. pyogenes* Type II CRISPR system is shown in Figure 1.9, demonstrating how CRISPR works in bacteria to protect the cell from infection by previously encountered foreign DNA. Viral DNA that enters the cell is recognised and cleaved into short 30bp 'protospacer' DNA fragments. This process is not fully understood, but Cas9 is known to play a crucial role, and

interestingly the HNH and RuvC nuclease domains of Cas9 are not required, suggesting that its role in creating protospacer DNA does not include cleavage[325]. Disruption of the PAM site recognition domain, which specifies the 'NGG' sequence required at the 3' of the target sequence for efficient cleavage, results in the creation of protospacer sequences with no PAM site specificity[326]. This suggests that the role of Cas9 at this stage is to identify eligible protospacer sequences upstream of potential PAM sites, to allow efficient cleavage when the same sequence is encountered upon re-infection. The requirement of a PAM site outside of the protospacer target sequence is an important feature that prevents Cas9 cleavage of the CRISPR array, which will contain an exact sequence match for each sgRNA produced.

While the mechanism for initial recognition of foreign DNA is not fully understood, presumably it is relatively inefficient, or there would be no requirement for the adaptive immune response element of CRISPR. It has been observed that protospacers matching the bacterial host genome are sometimes incorporated into CRISPR arrays, resulting in severe genomic instability[325]. This suggests that this initial recognition may not be efficiently targeted against foreign DNA, and would explain the necessity of acquired immunity, with cells targeting self undergoing negative selection, and cells targeting pathogens undergoing positive selection. This is similar to the selection processes applied to mammalian T-cells during development in the thymus[327].

After the recognition and cleavage of foreign DNA, Cas1, Cas2 and Csn2 are then required for the integration of the 30bp protospacer into the CRISPR array[325,326,328]. Cas1 forms a complex with Cas2, and has nuclease activity, able to target a specific sequence at the 5' end of the CRISPR array[328]. During this process, the new 30bp protospacer is inserted at the 5' end of the CRISPR array, and the 36bp repeating unit is duplicated, so that it flanks either side of the new protospacer.

The second phase of the CRISPR-Cas9 immune system, is the recognition and cleavage of foreign DNA upon reinfection. The entire CRISPR array is transcribed as one long crRNA, which is then broken down into sgRNAs, in a process coordinated by tracrRNA. tracrRNA have a 24bp region of sequence homology to the CRISPR array repeats, and recruit RNase III to these sites along the crRNA[316]. Cleavage occurs in both the protospacer sequence and the repeat region, leaving a sgRNA of 39-42bp, consisting of 20bp of protospacer target sequence at the 5', and 19-22bp of the repeat region at the 3' [316]. The subsequent sgRNA:tracrRNA complex binds to Cas9, and directs it to any sites matching the 20bp target sequence, which will be cleaved if a

PAM site (NGG) is situated immediately downstream of the target sequence. The secondary structure formed by the tracrRNA plays a key role in this process, and is required for Cas9 activity, both in identifying potential protospacer sequences in the first phase, and for the cleavage of them in the second phase[319,326].

1.7.3 CRISPR-Cas9 as a Laboratory Tool

For use in a laboratory setting, typically only the effector part of the CRISPR pathway is desired, and the Cas genes responsible for identifying novel foreign DNA elements, cleaving them and incorporating them into the host genome in the form of a CRISPR array, are omitted. The only components that are required are the sequence specific sgRNA, the tracrRNA and the effector complex. Typically, rather using two RNA components, the single fused gRNA model demonstrated by Doudna & Charpentier is used[319]. The most commonly used effector is based on Cas9, of the Class II Type II CRISPR System found in *S. pyogenes*.

1.7.3.1 Other CRISPR Systems and Cas9 Variations

There are two main classes of CRISPR system that have been identified in prokaryotes, defined by the component of the pathway responsible for binding to the tracrRNA, and introducing double strand breaks into the bound target DNA. Class I systems include Type I, Type III-A, Type III-B and Type IV CRISPR Systems, and rely on multiple subunits assembling to form one large complex, referred to as Cascade (**C**RISPR-**as**sociated **C**omplex for **A**ntiviral **De**fence) in Type I systems, Csm in Type III-A and Cmr in Type III-B[329]. The function and mechanism of Type IV systems are not fully understood, but have been identified based on sequence homology. Interestingly the Cas genes in Type IV are not always situated adjacent to a CRISPR array[329].

Class II includes Type II and Type V CRISPR Systems, these systems have one large gene that codes for the entire effector complex, that binds the tracrRNA and has the nuclease function, in Type II systems this is Cas9, and in Type V this is Cpf1[329].

Class II systems rather than Class I have been isolated and optimised for use in the laboratory, since it is much more efficient to clone and express a single gene than to dissect the Cas operon and clone several individual subunits. As was mentioned in 1.7.1, the initial four papers describing the use of CRISPR in human cells all used a Type II system based on Cas9 from *S. pyogenes*, optimised for human codon usage. This remains the most commonly used CRISPR

method in laboratories today, and various modified versions are now available, allowing researchers to tailor the Cas9 function to suit the requirements of their experiments.

Modified Cas9 variants are now available, including versions in which one of the nuclease domains is inactivated, either the HNH domain, which cleaves the strand bound by the gRNA or the RuvC domain, which cleaves the non-complementary strand[319]. The resultant Cas9 variants only cleave one strand, and are referred to as Cas9 'nickases', these only introduce double strand breaks if two gRNA are used in parallel on opposite strands, and greatly increase cleavage specificity, since if only one gRNA binds within a region, a nick rather than a double strand break is introduced, reducing off-target effects[330,331]. Other variants harbour mutations disrupting both nuclease domains; these $Cas9^{Null}$ variants do not have any nuclease activity, but can be tagged to another protein or domain with a regulatory function. In these cases CRISPR is used as a molecular delivery system, with Cas9-directed recruitment of a transcription factor or epigenetic modifier to a specific locus[332–334].

The Type V system is now also being used in laboratories, and is considered to have several advantages over the Type II system. Type V is still a Class II System, and so still only requires a single molecule effector (Cpf1), and Cpf1 is smaller than Cas9, and can therefore be introduced into cells more easily. Endogenous Cpf1 from *Francisella novicida* is 3,900bp, while Cas9 from the same strain (*F. novicida* has both systems) is 4,887bp, and the commercially available optimised Cas9 from *S. pyogenes* is 4,101bp[335]. The sgRNA produced after processing of the crRNA in the Type V system interact directly with Cpf1, and no tracrRNA is required, similarly to the gRNA system designed by Doudna & Charpentier for the Type II system[319,335]. The PAM site required for cutting by Cpf1 is 'TTN' at the 5' of the target sequence[335]. The fact that this is different from the 'NGG' favoured by Cas9 provides additional flexibility in gRNA design, which is limited by presence of a PAM site in the target region, e.g. if the region of interest has a low abundance of 'GG' dinucleotide, the Cpf1 system can be used instead (and vice versa). Additionally, while Cas9 produces blunt ends, Cpf1 cleavage leaves a 5-nucleotide overhang at the 5' end, this can be used for efficient template insertion if an artificial template with the same overhangs is generated[335]. Since the cleavage sites are not palindromic like the sequences typically targeted by restriction endonucleases, the overhangs produced are not identical (unless by chance) and allow directional insertion of the template sequence.

1.7.3.2 Endogenous Repair Machinery

While the gRNA directed cleavage of DNA by Cas9 is highly specific and efficient, the subsequent Double Strand Break (DSB) repair relies on the host's endogenous repair machinery, and is less predictable.

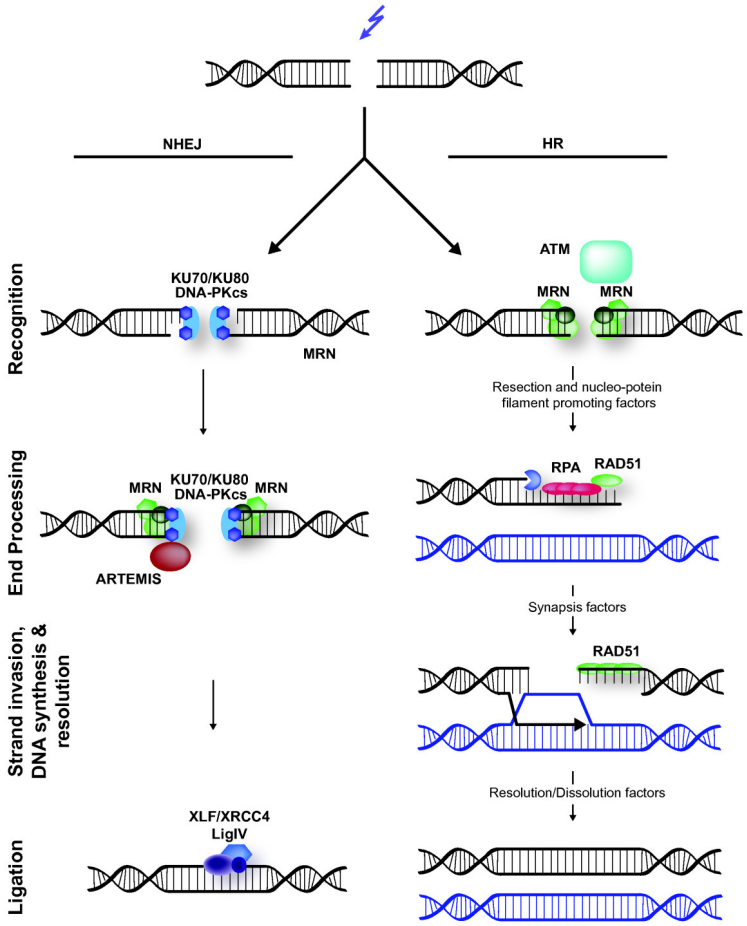


Figure 1.10: Overview of the two main DSB repair pathways in humans. NHEJ – Non Homologous End Joining. HR – Homology Directed Repair. NHEJ involves the identification of DSB ends by Ku70/80, followed by non-specific end processing and ligation by Ligase IV. HDR pathway uses homologous sequence as a repair template to correct the damaged sequence. Image from Lans *et al.* (2012)[336].

Two main pathways repair DSBs in humans, Non-Homologous End Joining (NHEJ) or Homology Directed Repair (HDR), as summarised in Figure 1.10. NHEJ is non-specific, with ku70/80 recognising two dsDNA ends, and Ligase IV ligating them, often resulting in introduction of short insertions or deletions.

HDR relies on the identification of a homologous sequence to use as a template to repair the gap between the two ends, which would usually be the other allele. In HDR, the MRN complex recognises and binds to the DSBs, and recruits other factors including StIP, which is required for trimming the 5' strand, leaving a 3' overhang at the DSB site[337,338]. RAD51 displaces the bound RPA, and enables the 3' overhang to 'invade' nearby dsDNA. When a region of

sequence homology is found, the overhang acts as a primer, and polymerases extend the sequence, reading off the template strand[338]. This continues until the junction between the two sister chromosomes is resolved, which can occur by a variety of different methods, and may result in recombination and genomic crossover[338].

For gene knockouts or introduction of non-specific deletions to disrupt a DNA motif, the NHEJ pathway is sufficient, and produces a variety of different genotypes with differing lengths of insertions or deletions as a result of end processing. However, for precise gene editing, where a specific sequence is desired, the HDR pathway is relied on. Under normal conditions, the HDR pathway uses the undamaged allele as the template for repair, but an artificial DNA repair template can be provided, containing a modified sequence for incorporation into the genome. Artificial DNA repair templates can either be incorporated into a plasmid, or introduced as independent DNA fragments, which can be either double stranded or single stranded[339,340].

Utilisation of HDR as a mechanism to incorporate template DNA into the genome after introduction of DSB is well established[341,342]. This was initially developed through the use of older techniques such as meganucleases, which bind and cleave specifically at long target sequences, which could be modified for sequences of interest[342,343]. More recently, this has developed into the use of fusion proteins, such as Zinc Finger Nuclease (ZFNs) and Transcription Activator Like Effector Nucleases TALENs, fusing nuclease domains to DNA binding domains with high sequence specificity, essentially allowing modular construction of a nuclease that targets a desired site[342,344–346]. This is very similar to the principle of CRISPR-Cas9 targeted cleavage, but the ease with which gRNA can be substituted enables Cas9 based approaches to be more efficiently applied in practice, compared to fusion proteins which require construction of a new specific protein for each target site[342].

Template DNA for HDR can be either single or double stranded, and can be introduced to the cell directly, or incorporated into a vector such as a plasmid or virus[347–349]. When introducing the template sequence in the form of Single Stranded Oligodeoxynucleotides (ssODNs), efficient incorporation can be observed with sequence homology arms of >40bp flanking the targeted DSB site[347]. However, when using dsDNA, longer homology arms are required, typically ranging from 500-800bp either side of the target site, and linearised DNA is more efficient than a circular plasmid[340,350,351].

In order to improve the efficiency of cleavage directed genomic editing techniques such as CRIPSR, which rely on the incorporation of a template sequence into the host genome, the

HDR pathway is preferred over NHEJ. There are several techniques that have been suggested as a way to promote HDR over NHEJ in cells, including both transitory changes, such as siRNA knockdown of NHEJ components, as well as permanent changes, such as gene knock outs[352–354]. While knock outs are more efficient, and would effectively silence the NHEJ pathway, they risk substantial genomic instability in any subsequent cell lines generated[354–356]. Transitory silencing is therefore preferable, since the loss of NHEJ function can be limited to a short time frame during which genome editing is taking place.

Alternative methods to promote rates of HDR over NHEJ include small molecule inhibitors of the NHEJ pathway, such as the Ligase IV inhibitor SCR7[357,358], as well as taking advantage of the differences in activity observed during different phases of the cell cycle. NHEJ is active throughout the cell cycle, whereas HDR activity is mostly active during S-Phase[359]. It has recently been shown that the use of modulators of the cell cycle, such as Cyclin D1, can increase HDR efficiency by promoting transition to S-Phase[351]. Modified versions of Cas9 fused to a Geminin, a protein that is targeted for degradation during G1 have also been shown to increase efficiency, allowing temporal regulation of Cas9 machinery to ensure that cleavage occurs at a time when HDR is most efficient[360,361].

1.7.3.3 Off-Target Activity of CRISPR-Cas9

The off-target effects of CRISPR-Cas9 directed cleavage of DNA are well documented. ChIP-Seq experiments have demonstrated Cas9 occupancy at off-target sites, and while this is believed to overestimate off-target effects, since Cas9 occupancy does not necessarily correlate with actual cleavage, whole genome sequencing analyses identify a large number of indel mutations and SNPs at loci across the genome, in both coding and non-coding regions[362–366]. This off-target activity is mostly unpredictable, and could severely restrict the conclusions that are able to be drawn from experiments replicating mutations using this technique. To verify that any observed effects are caused by the mutation of interest rather than secondary mutations introduced at other genomic loci, it may be necessary to carry out costly genomewide analyses to document all the specific off-target effects for each clonal cell population generated.

The majority of off-target activity appears to arise as a result of *S. pyogenes* Cas9 tolerating some mismatches between the gRNA and the bound DNA, with increased tolerance further

away from the PAM site, and frequency of cleavage at these off target sites also increases at higher Cas9 concentrations[367,368].

While the off target effects can be mediated to a certain extent by selecting gRNA with the fewest predicted off-target sites, based on sequence similarity in the target host genome, there is only limited scope for this, especially when introducing specific mutations, since the gRNA need to be as close to the target site as possible to maximise efficiency for HDR. Instead, specificity may be improved through the use of other Cas9 variants, such as the Cas9 nickases described in .1.7.3.1, where two different Cas9 molecules must bind on opposite strands and in close proximity to cleave DNA[324]. Modified versions of Cas9 have now also been generated specifically for their improved specificity and reduction in off-target activity, these include High Fidelity Cas9 (spCas9-HF1), Enhanced Specificity Cas9 (eSpCas9) & Hyper Accurate Cas9 (HypaCas9)[369–371]. Interestingly, truncated gRNA (tru-gRNA) with target sequences shorter than the usual 20bp have also been demonstrated to reduce Cas9 off-target activity[372].

1.7.4 Current Clinical Work

The development of accurate genomic editing techniques such as CRISPR-Cas9 present many therapeutic opportunities, particularly for monogenic blood disorders such as SCA. The sickle cell mutation has already been corrected using CRISPR-Cas9 in human erythroid progenitors *in vitro*, although with a very low efficiency, highlighting the issue of relying on the endogenous HDR machinery[373].

An alternative strategy has been to introduce large scale deletions at the β-globin locus, mimicking a naturally occurring 13kb deletion that causes HPFH, increasing γ-globin expression[374]. Introducing genomic deletions relies on the NHEJ pathway rather than HDR and is much more efficient, correction of the sickle mutation requires a highly specific gene-edited product, whereas the lack of specificity of the large deletion allows for more sequence variation surrounding the deletion site after the DSB repair.

Chapter 2 Materials & Methods

2.1 Isolation of Erythroid Progenitors

2.1.1 Blood Samples & PBMC Isolation

Peripheral blood samples were collected from either healthy donors, or anonymised SCA patients, recruited from Guy's Hospital or King's College Hospital to participate in projects IRAS ID - 128238 or IRAS ID - 35853.

9 – 27ml blood were collected in vacuum tubes containing EDTA. Samples were centrifuged at 3000rpm for 10 minutes, and the Buffy Layer collected using a Pasteur pipette. The Buffy Layer was diluted 1:1 in Dulbecco's Phosphate-Buffered Saline (PBS), gently layered onto 2/3 by volume of pre-warmed Histopaque®-1077 (Sigma-Aldrich – 10771) in polystyrene culture tubes (Corning - 430172). Samples were centrifuged at 1900rpm for 30min with no brake, and the PBMC layer was collected using a Pasteur pipette.

PBMC samples were then washed three times in PBS before progressing to downstream steps.

2.1.2 Culture Conditions

Phase 1: PBMCs were plated at 10^7 cells per ml, in StemSpan™ media (Stem Cell Technologies - 09650), supplemented with 1 nM Dexamethasone (Sigma-Aldrich – D4902), 2 U/ml Erythropoietin (Eprex), 40 ng/µl Insulin-like Growth Factor 1 (R&D Systems – 291-G1-200), 2 ng/ml Interleukin-3 (Stem Cell Technologies - 02603), 2 mM L-Glutamine (Sigma-Aldrich – G7513), 1% Penicillin & Streptomycin (Sigma-Aldrich – P4333), 40 ng/µl Stem Cell Factor (Sigma – S7901) and 0.2% Synthechol (Sigma-Aldrich – S5442). The cultures were transferred to new plates daily, in order to remove adherent cells. The medium was refreshed on Phase 1 Day 3 (P1D3), when cells were diluted to 2 x 10^6 cells per ml, and Interleukin-3 concentration was reduced to 1 ng/ml.

Phase 2: Phase two was initiated on either P1D6 or P1D7, depending on the health of the culture as assessed by cytospin. The medium was fully replaced, without Interleukin-3, and cells were diluted to 10^6 cells per ml, and maintained at this concentration for the remainder of the culture. Throughout Phase 2, cultures were monitored daily by cytospin.

2.1.3 Cytospins

Aliquots were concentrated onto microscope slide by centrifugation at 1000rpm for 3min in a cytocentrifuge (Thermo Scientific™ Cytospin™ 4). Slides were incubated for one minute in each of methanol fixation solution, eosin solution, and methylene blue solution (Thermo Fisher Scientific - 10435310). Stained cells were air-dried before addition of DPX Mountant and cover slip.

2.1.4 Flow Cytometry & Cell Sorting

Cells were re-suspended in 80µl PBS and incubated for 20min at room temperature with 20µl Human Fc Receptor Binding Inhibitor (eBioscience - 14-9161-73). 5µl of antibody was added, and cells were incubated for 30min at room temperature in the dark and washed with PBS prior to analysis or sorting.

Flow cytometry analysis was carried out using BD Accuri™, BD FACSCalibur™ or BD FACSCanto II™ cytometers. Fluorescence-Activated Cell Sorting (FACS) was carried out on BD FACSAria™ cell sorters, as a service provided by the BRC Flow Cytometry Core Facility at Guy's Hospital. The FACS service was unavailable outside of working hours, and so for direct sorting of SCA patient PBMCs, cells were plated in the culture medium overnight before sorting the following day, at P1D1.

Fluorescent antibodies used for flow cytometry: anti-CD71-APC (eBioscience - 17-0719-42), anti-Glycophorin A-FITC (eBioscience - 11-9886-42), anti-CD45-eFluor® 450 (eBioscience - 48-9459-42), anti-CD34-PE (Miltenyi Biotec - 130-098-140), anti-CD14-PE-Cy7 (eBioscience - 9025-0149-120) and anti-c-Kit-PE-Cy7 (eBioscience - 25-1178-42).

2.1.5 Antibody-MicroBead Cell Isolation

For $CD71^+$ cell isolation, freshly extracted PBMCs were depleted of $CD45^+$ cells using Human CD45 MicroBead kit (Miltenyi Biotec - 130-045-801) using LD columns. The CD45- fraction was then enriched for $CD71^+$ cells using Human CD71 MicroBead Kit (Miltenyi Biotec - 130-046-201) using MS columns.

$CD34^+$ cells were isolated from freshly extracted PBMCs using Human CD34 MicroBead Kit UltraPure (Miltenyi Biotec - 130-100-453) using MS columns, the protocol was repeated with an additional column, for increased purity.

For isolation of $GPA^{-}CD71^{+}$ cells, freshly extracted PBMCs were depleted of GPA^{+} cells using Human CD235a (Glycophorin A) MicroBead Kit (Miltenyi Biotec – 130-050-501) using LD columns. The GPA- fraction was then enriched for $CD71^{+}$ cells using Human CD71 MicroBead Kit (Miltenyi Biotec - 130-046-201) using MS columns.

2.1.6 DNA & RNA extractions

DNA & RNA were extracted simultaneously from isolated erythroid progenitor cells either after storage in TRIzol® Reagent (Thermo Fisher Scientific - 15596018) at -80°C, or immediately after isolation using a Qiagen AllPrep DNA/RNA/Protein Mini Kit (Qiagen - 80004), in conjunction with a QIAshredder (Qiagen - 79654) for homogenisation. Where DNA & RNA were extracted from cells separately, DNA was extracted using a Qiagen DNeasy Blood & Tissue Kit (Qiagen - 69504), Qiagen Puregene Blood Core Kit A (Qiagen - 158445), or a Qiagen QiaAMP DNA Micro Kit (Qiagen - 56304), and RNA was extracted using a Qiagen RNeasy Mini Kit (Qiagen - 74104). Concentrations were analysed by NanoDrop Spectrophotometer (NanoDrop 2000 or NanoDrop One) or by Qubit using Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific - Q32854) and Qubit® RNA HS Assay Kit (Thermo Fisher Scientific - Q32852) for DNA and RNA respectively.

2.2 Whole Exome Sequencing

2.2.1 SCA Patient WES Data

2.2.1.1 SCA Patients from King's College Hospital

Samples were selected from a collection of >700 SCA patients managed by Professor Swee Lay Thein and her team at King's College Hospital. These patients had been recruited for participation in genetic research, under projects LREC 01-083, 07/H0606/165 or 12/LO/1610. Patients were classified as either mild or severe in consultation with Professor Thein & Dr Catherine Gardner.

Genomic DNA (gDNA) samples of >1.5µg were submitted to the NIHR Biomedical Research Centre Genomics Core Facility at Guy's and St Thomas' NHS Foundation Trust, for Whole Exome Sequencing (WES) library preparation and sequencing. Exome capture was carried out using Agilent SureSelectXT Human All Exon v5 kit (Agilent – 5190-6210), and samples were sequenced on an Illumina HiSeq2000. Sequence reads were mapped to the reference genome (GRCh37/hg19), assessed for read quality and variant calls were annotated, this was carried out using an in-house analytical pipeline developed by Professor Michael Simpson.

2.2.1.2 SCA Data from dbGaP Dataset

Exome sequencing data from clinical trials investigating response to HU therapy in SCA patients were obtained through dbGaP (Study Accession ID: phs000691.v2.p1)[375]. Analyses of the data generated in this study were initially published by Sheehan *et al.* in 2014[376].

These samples were sequenced in the Human Genome Sequencing Center at Baylor College of Medicine, exome capture was carried out using Roche SeqCap EZ HGSC VCRome 2.1 kit (Roche NimbleGen - 06465587001), and samples were sequenced on an Illumina HiSeq2000.

651 patients were included in the study data, of which 143 were recruited as part of the Long Term Effects of Hydroxyurea Therapy in Children With Sickle Cell Disease clinical trial (HUSTLE - NCT00305175), 132 of the Stroke With Transfusions Changing to Hydroxyurea clinical trial (SWiTCH - NCT00122980) and 139 of the Transcranial Doppler (TCD) With Transfusions Changing to Hydroxyurea clinical trial (TWiTCH - NCT01425307), the remaining 237 had no annotated source.

Short read archive files were downloaded from dbGaP (phs000691.v2.p1), converted into FASTQ format, and then processed through the same alignment, quality control and variant

calling pipeline as the in-house exomes, in order to minimise variation between the two data sources.

2.2.2 Filtering of ANNOVAR Annotated Variants & Statistical Testing

Bioinformatic tools were used to identify relevant candidate variants, and to perform statistical tests for significance between the mild and severe SCA patient groups.

2.2.2.1 Computational Tools for Filtering

Details of the individual variant filtering criteria used for each specific analysis performed is described in full in 4.3.1. Variant filtering steps were performed on the British Research Council's Athena-Apollo High Performance Computing Cluster at KCL, using Python v2.7.12.

2.2.2.2 Fisher's Exact Test

Fisher's exact test was performed using the SciPy Python for Scientific Computing Toolkit[377].

2.2.2.3 CADD Phred-Like Variant Scoring

Combined Annotation Dependant Depletion (CADD) Phred-like scores were obtained for coding variants and loss of function mutations from http://cadd.gs.washington.edu[378].

2.2.3 Manual Assessment of Variants

Top candidate variants were assessed individually based on the proximity of the variant to any annotated structures within the gene, and their inclusion in any alternative splicing isoforms.

2.2.3.1 Identification of Gene Features

Information regarding annotated protein domains and conserved regions as well as structural information was accessed from the neXtProt knowledgebase on human proteins[379].

2.2.3.2 Identification of Alternative Isoforms

Alternative splicing isoforms were investigated using the University of California Santa Cruz (UCSC) Human Genome Browser[380].

2.3 CRSIPR-Cas9 Plasmid

CRISPR-Cas9 plasmids were provided by Horizon Discovery Group, as part of a programme to test their gUIDEbook™ gRNA design platform set up through a partnership with Desktop Genetics Ltd. The plasmids provided were derived from pD1301, and key features include a kanamycin resistance gene, a chimeric gRNA scaffold and a Cas9 tagged with a self-cleavable linker to DasherGFP.

The gRNA consists of a 20bp target sequence, identical to the region of interest in the host genome, coupled to a 76bp gRNA scaffold that forms a secondary structure to interact with Cas9. Expression of the gRNA is controlled by the P_hU6.1-human RNA expression promoter containing a TATA box.

The Cas9 is derived from that found in *Streptococcus pyogenes*, and has nuclease activity capable of introducing double strand breaks in mammalian DNA, directed by interactions with the gRNA. Expression of Cas9 is controlled by E_CMV, a cytomegalovirus enhancer element, and P_CMV, a constitutive mammalian promoter with strong expression. Cas9 is tagged at both the 5' and 3' ends with nuclear localisation signals (NLS), which direct proteins for nuclear import, ends. Downstream of the 3'NLS is a CHYSEL_TAV linker connecting to DasherGFP. CHYSEL (cis-acting hydrolase element) will self-cleave in the cytoplasm after translation, leaving cytosolic GFP as a marker for expression, while Cas9 is targeted to the nucleus. Downstream of DasherGFP is pA_GH-bovine(min), a polyadenylation signal.

The kanamycin resistance gene encodes Neomycin phosphotransferase II and allows survival in medium containing kanamycin. Expression is controlled by bacterial promoter P_Amp.

Ori_pUC is an origin of replication sequence derived from *E. coli*, allowing the plasmid to replicate when transformed into competent bacteria.

Bacterial transcriptional termination cites Term_rpoC & Term_bla are included upstream of the human expression promoter, to prevent read through and unregulated expression of gRNA & Cas9 in bacterial cells.

The full plasmid map is shown in Figure 2.1 and the full sequence is provided in Appendix 2, with the 20bp variable gRNA target sequence highlighted.

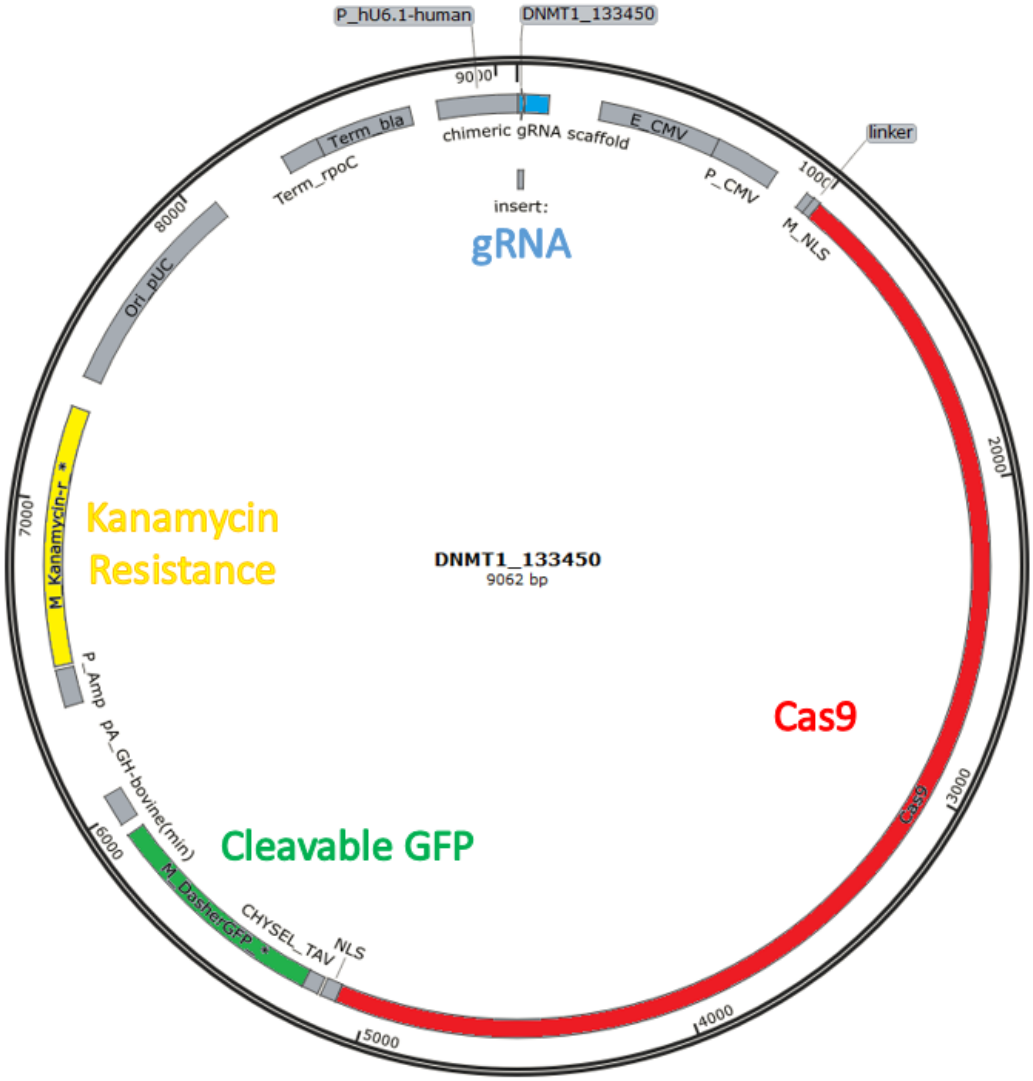


Figure 2.1: Plasmid map of the 9kb pD1301 Cas9 plasmid provided by Horizon Discovery Group. Key features are highlighted: Cas9 is shown in red, self-cleaving GFP tag in green, kanamycin resistance gene in yellow, and gRNA target sequence and scaffold shown in blue.

2.4 CRIPSR-Cas9 Genomic Editing

2.4.1 gRNA Design

gRNA to target the SNP sites in the ASH1L and KLF1 genes were designed using the online tool DESKGEN[381]. Five candidate gRNA were initially selected for each SNP based on proximity of the cleavage site to the SNP, and the 'on-target score', a scoring mechanism designed to estimate the likelihood of Cas9 cleavage[382,383].

					PAM Site		*Disrupted PAM Site*	
	gRNA	*Off Target*	*On Target*	*Distance*	*Protein*	*DNA*	*Protein*	*DNA*
K1	GATCTCAGCT TAGTCTGGCA	89	62	22bp	n/a	G**GG**	n/a	G**C**G
K2	AGGTACGCTC AGTCCAGGAG	87	61	3bp	n/a	A**GG**	n/a	A**GC**
K3	AGTCTGGCAG GGGGTGAGGA	50	51	34bp	n/a	G**GG**	n/a	G**C**G
K4	TAAGCTGAGA TCTCCTCTCC	79	51	1bp	n/a	T**GG**	n/a	T**GC**
K5	AAGAGACTTA ACCAGGACTG	73	68	23bp	n/a	A**GG**	n/a	A**C**G
A1	TCTTCCGGCC ACTGGAGTTA	85	45	17bp	NP	AA**C** **C**CT	NP	AA**T** **C**CT
A2	CAAACCCTAA CTCCAGTGGC	88	57	24bp	R	C**GG**	R	C**GA**
A3	GTTTCAAACC CTAACTCCAG	61	70	20bp	G	**GG**C	Not Possible	-
A4	AACCTTTTCAC AAGTGCAAT	77	50	14bp	G	**GG**C	Not Possible	-
A5	GTATGTTCATC ACTGCTGGC	83	46	49bp	P	**CC**A	Not Possible	-

Table 2.1: Five candidate gRNAs for both the KLF1 (K1-5) and ASH1L (A1-5) SNPs. The on-target and off-target scores are shown, along with the gRNA sequence and distance between the target SNP and the cleavage site. The codons in which the endogenous PAM sites are situated are shown, with the GG dinucleotide in bold. Red indicates proposed changes to disrupt the PAM site. K1 & K2 were selected for KLF1, due to high off-target scores, which were considered more important. A1 & A2 were selected for ASH1L, since they were the only gRNAs with PAM sites that could be silently disrupted, they also have high off-target scores.

The five gRNAs are shown in Table 2.1, of these five, two gRNA were selected for use for each SNP, these were chosen based on the 'off-target score', an inverse scoring mechanism to predict the likelihood of Cas9 cleaving non-targeted regions of the host genome with a similar sequence to the gRNA[368]. The ability to introduce a silent mutation to disrupt the PAM site was also investigated at this stage. Due to the fact that the KLF1 SNP is intronic, and thought to disrupt transcription factor binding, any disruption of the PAM site may also influence

transcription factor binding, making it difficult to distinguish which of the two mutations is the cause of any effect observed. Therefore for the KLF1 SNP, a negative control was carried out, introducing only the PAM mutation.

2.4.2 Plasmid Design & Cloning

A DNA sequence containing the target SNP and the PAM site disruption was cloned into each plasmid. This sequence acts as a template, and is incorporated into the host by Homology Directed Repair (HDR) after cleavage of the genomic DNA target sequence by Cas9. Including the PAM site disruption in the template sequence not only prevents Cas9 from repeatedly cutting the genomic DNA once the sequence has been incorporated, but also prevents cleavage of the template DNA in the plasmid.

Initially, unmodified template DNA (containing the wild type sequence) was cloned into each plasmid. A 20bp gRNA target sequence was then substituted in by Site Directed Mutagenesis (SDM). A PAM site disruption mutation was then introduced to the template DNA in each of these plasmids, specific to the gRNA target sequence. This was also achieved by SDM. Aliquots of plasmids at this stage were kept for use as PAM site only controls. The SNPs of interest were then introduced to the template DNA, also by SDM. This process is outlined in Figure 2.2.

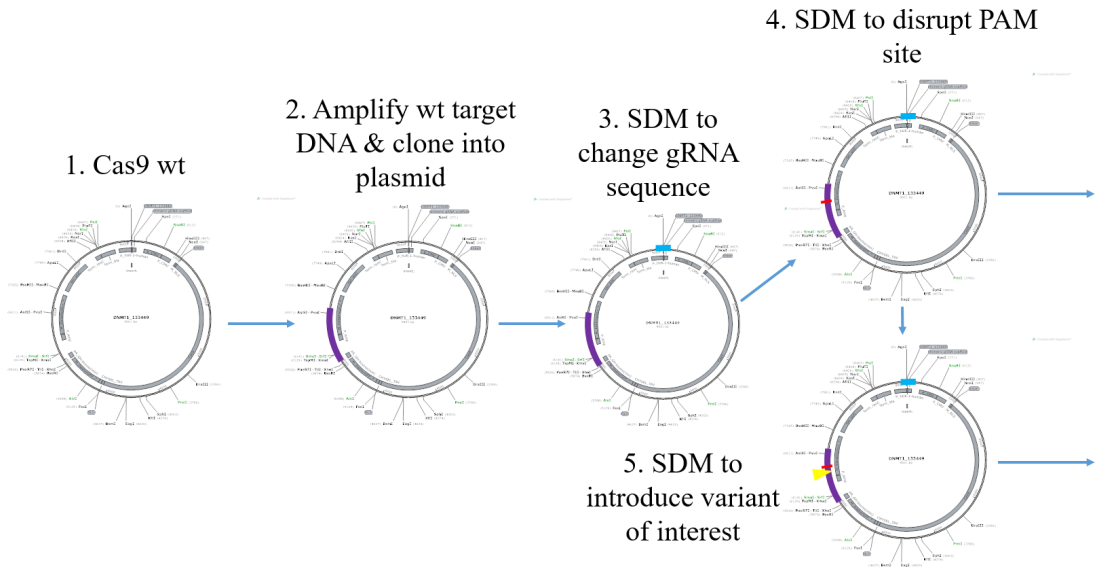


Figure 2.2: Diagram showing the cloning workflow to generate plasmids for introduction of specific genetic variants using the CRISPR-Cas9 system. Template sequence is indicated in purple, gRNA in blue, PAM site disruption in red, and SNP in yellow.

2.4.2.1 Unmodified Template DNA Insertion

Template DNA was inserted into the CRISPR-Cas9 plasmids at a BssHII restriction site, situated between the kanamycin resistance gene and the origin of replication sequence. Unmodified template DNA was amplified from K562 genomic DNA by PCR, using primers tagged with a 6bp spacer and a 6bp BssHII restriction site at the 5' end. PCR products were Sanger sequenced to confirm that the K562 sequence matched that of the reference genome, since cancer cell lines typically carry high levels of genetic variation. Sequences of the primers used for amplification are shown in Table 2.2 and the KLF1 & ASH1L regions amplified are shown in Figure 2.3. Full gene maps are provided in Appendix 1.

The whole 20µl PCR reaction was run on an agarose gel, and bands were excised and DNA extracted as described in 2.5.4.2. Plasmids and PCR products were digested with BssHII (NEB – R0199S). 1µg plasmid, or up to 1µg of PCR product were added to 5µl 10X NEB CutSmart® Buffer and 1µl BssHII, and made up to a total reaction volume of 50µl with nuclease free H_2O. Digests were run at 50°C for 1 hour, and then 65°C for 20 minutes to heat inactivate the enzyme.

Products of the 50µl digest were run on an agarose gel, excised and extracted. Ligation was then performed using T4 DNA Ligase (NEB – M202S). 50ng of digested and purified plasmid was added to 2µl 10X T4 DNA Ligase Buffer and 1µl T4 DNA Ligase, the remainder of the 20µl reaction volume was then made up with the digested and purified PCR product. The reaction was incubated at 4°C overnight.

10µl of the ligation reaction was transformed into competent cells and plated on agar with kanamycin (as described in 2.5.8). Colonies were screened for successful template insertion by colony PCR and Sanger sequencing.

Primer	*Sequence*
KLF1_Temp1_F	TAAGCA**GCGCGC**CCGCTGATATCTGGAAGATTGT
KLF1_Temp2_R	TAAGCA**GCGCGC**CTTGCCTTGCTTTGCCTTATC
ASH1L_Temp2_F	TAAGCA**GCGCGC**CCTGCATACTACTAACAGACCTATG
ASH1L_Temp1_R	TAAGCA**GCGCGC**TGCTGTTGGAGAGCGATATAAG

Table 2.2: Table showing primer sequences used for amplification of KLF1 & ASH1L template DNA. GCGCGC BssHII restriction site is shown in red and bold, 6bp spacer at 5' of restriction site is shown in blue.

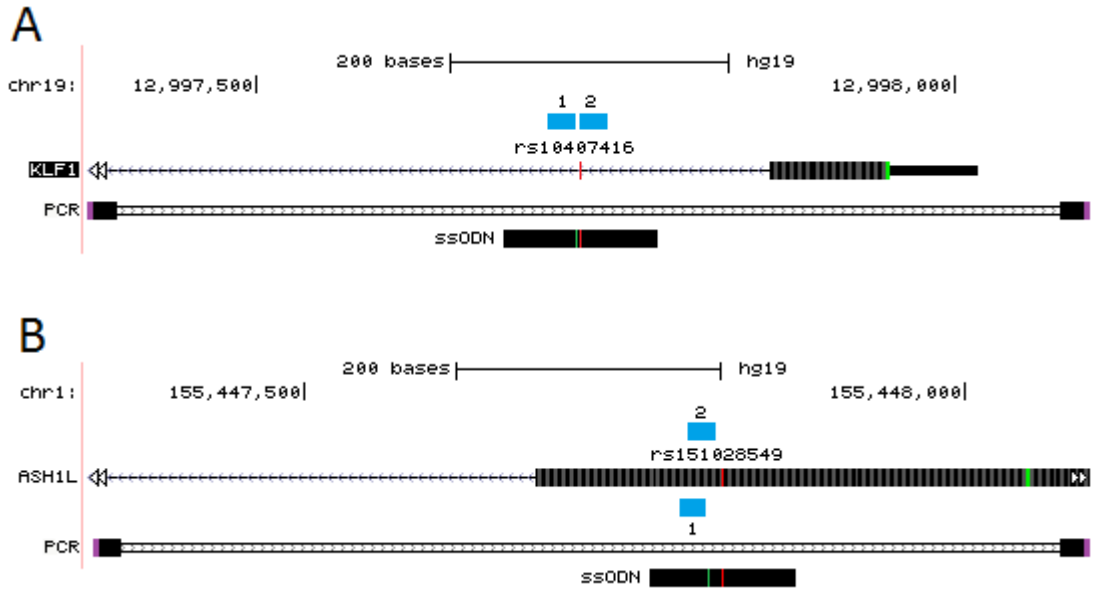


Figure 2.3: Diagrams of PCR amplicons used to clone K562 genomic DNA into CRISPR-Cas9 plasmids to act as a template for Homology Directed Repair (HDR). Images are adapted from UCSC Genome Browser (*http://genome.ucsc.edu* - Assembly GRCh37/hg19[380]). A – 718bp amplicon from KLF1. B – 759bp amplicon from ASH1L. PCR amplicons are shown below the genomic sequence, with BssHII restriction site tags at 5' of primer in purple. In the genomic DNA sequence the targeted SNPs are indicated by red lines, with methionine residues and start codons indicated in green. gRNA target sequences are highlighted in blue. Also shown are single stranded oligodeoxynucleotides (ssODN), which were designed as an alternative technique to introduce the template sequence. In the ssODNs the SNP is shown in red, and the PAM site disruption in green. Full gene maps are shown in Appendix 1.

2.4.2.2 gRNA Sequence Substitution

gRNA sequences were introduced to plasmids with successful template insertion. This was done using Site Directed Mutagenesis (SDM), performing a substitution of the initial 20bp gRNA sequence for one of those designed in 2.4.1. Primer sequences used for SDM are shown in Appendix 3, the 10bp at the 5' of both forward and reverse primers contain the variable gRNA sequence, and the 3' ends anneal to the sequences flanking the insertion site. Therefore all forward primers follow the pattern NNNNNNNNNN-GTTTTAGAGCTAGAAATAGCAAG, where N_{10} represents the last 10bp of the target sequence, and all reverse primers follow the pattern NNNNNNNNNN-CGGTGTTTCGTCCTTTCC, where N_{10} represents the reverse complement of the first 10bp of the target sequence.

After the SDM reaction and plating of transformed DH5α on kanamycin plates (as described in 2.5.6), colonies were screened by Sanger sequencing for incorporation of the new sequence. Primers spanning the gRNA site are shown in Appendix 3.

2.4.2.3 PAM Site Disruption & SNP Introduction in Template Sequence

Plasmids containing the template insertion and the gRNA substitution underwent sequential rounds of SDM, transformation and cloning to introduce firstly a PAM site disruption mutation, as designed for each gRNA and shown in Table 2.1, and secondly the SNP of interest. The sequences for the SDM primers are shown in Table 2.3.

Primer	Sequence
PAM Site Disruption	
A1_PAMF	TGGAGTTAGGATTTGAAACTCTG
A1_PAMR	GTGGCCGGAAGAAATTAAC
A2_PAMF	CCAGTGGCCGTAAGAAATTAAC
A2_PAMR	AGTTAGGGTTTGAAACTCTG
K1_PAMF	AGTCTGGCAGCGGGTGAGGAG
K1_PAMR	AAGCTGAGATCTCCTCTCC
K2_PAMF	TCCAGGAGAGCAGATCTCAGC
K2_PAMR	CTGAGCGTACCTCAGTCC
SNP Introduction	
A1_SNPF	TTTGAAACTCCGCAGCTGCCTATTG
A1_SNPR	TCCTAACTCCAGTGGCCG
A2_SNPF	TTTGAAACTCCGCAGCTGCCTATTG
A2_SNPR	CCCTAACTCCAGTGGCCG
K1_SNPF	GAGATCTCCTGTCCTGGACTGAG
K1_SNPR	AGCTTAGTCTGGCAGCGG
K2_SNPF	GAGATCTGCTGTCCTGGACTGAG
K2_SNPR	AGCTTAGTCTGGCAGGGG

Table 2.3: SDM primer sequences for PAM site disruption and SNP introduction to the template sequence in the CRISPR-Cas9 plasmid. PAM site disruption SNPs are highlighted in green, with targeted SNPs in red. In cases where the gRNA target sequence is close to the SNP, the PAM site is also close, in these cases both the PAM site disruption and the SNP must be included in the second SDM reaction, to prevent the SNP introduction SDM reversing the PAM site disruption. The possibility of using a single SDM reaction to introduce both variants was considered for these cases, this was rejected since it would not enable production of PAM only controls.

2.4.3 siRNA Knock Down of Non-Homologous End Joining Pathway

The first and last components of the NHEJ pathway, XRCC6 and LIG4, were targeted for knock down using siRNA.

siRNA were ordered for both XRCC6 (OriGene - SR301689) and LIG4 (OriGene - SR302689), with three separate siRNAs provided for each target gene, as well as one vial of Universal Scrambled siRNA as a negative control (OriGene – SR30004). siRNAs were 27 ribonucleotides long, and were provided as 2nmol of lyophilised powder, which were resuspended in 100µl of RNAse free siRNA duplex resuspension buffer (OriGene – SR30005) to a final concentration of

20μM, and incubated at 94°C for 2 minutes before first use. Sequences of the siRNA, as well as the primers used for rtPCR analysis of knockdowns are shown in Appendix 4.

siRNA were transfected alongside CRISPR plasmids by nucleofection, with 1.5μl (30pmol) of each of the three siRNA for either target gene, or 4.5μl (90pmol) scrambled siRNA, and up to 5μg of CRISPR-Cas9 plasmid.

2.4.4 Single Stranded Oligodeoxynucleotide (ssODN) Templates

As an alternative to the template sequence incorporated into the plasmid, 110bp ssODNs with 55bp sequence homology either side of the SNP were designed, and ordered from Eurofins Genomics. The ssODN templates also contained the PAM site disruption. This was done for gRNA A2 for ASH1L, and gRNA K3 with the corresponding PAM only template, for KLF1. Templates were designed antisense to transcription, since this has been shown to increase incorporation into the genome[384]. The ssODN sequences are shown in Table 2.4. ssODN alignments for gRNAs A2 and K3 against the ASH1L gene and KLF1 gene respectively are shown in Figure 2.3. ssODNs were transfected alongside the plasmids containing the relevant gRNA by nucleofection, with 30μg ssODN and up to 5μg plasmid.

ssODN Target Gene & SNPs	ssODN Sequence
ASH1L - A2	AGTCCAGGGCTGTCAGTTAATTTCTTCCGGCCACTGGAGTTAGG**A**TT TGAAACTC**G**GCAGCTGCCTATTGCACTTGTGAAAAGGTTTGTATGTTC ATCACTGCTGGCTGG
KLF1 - K3	CTCAAACCCCTAGACCACCCTCCTCACCCCCTGCCAGACTAAGCTGA GATCT**G**CT**G**TCCTGGACTGAGCGTACCTCAGTCCTGGTTAAGTCTCT TGATTTCAGGTCAAGA
KLF1 - K3 PAM Only	CTCAAACCCCTAGACCACCCTCCTCACCCCCTGCCAGACTAAGCTGA GATCT**G**CT**C**TCCTGGACTGAGCGTACCTCAGTCCTGGTTAAGTCTCT TGATTTCAGGTCAAGA

Table 2.4: Sequences for 110bp ssODN templates used. PAM only control was used in parallel for KLF1 but not ASH1L, due to the fact that the PAM disruption is translationally silent. PAM disruptions are highlighted in green, with the targeted SNPs in red.

2.5 Molecular Biology & Cloning Tools

2.5.1 Oligonucleotide Primers

Oligonucleotide primers for PCR, Sanger sequencing and rtPCR were designed using Primer3Plus Version 2.4.0[385]. Primers for Site Directed Mutagenesis were designed using the online tool NEBaseChanger™[386]. All primers were ordered from Eurofins Genomics.

2.5.2 Polymerase Chain Reaction (PCR)

Unless otherwise specified, PCR was performed using ThermoPrime 2X ReddyMix PCR Master Mix (Thermo Scientific - AB0575DCLDB), and run on an MJ Research PTC-200 thermal cycler. The standard reaction mix for a 20µl reaction, and the standard PCR programme are shown below in Table 2.5.

PCR Reaction Mix

Reagent	*Volume*
2X ReddyMix	10µl
Forward Primer 20µM	0.5µl
Reverse Primer 20µM	0.5µl
Nuclease Free H_2O	8µl
Template DNA	1µl
Total	***20µl***

PCR Thermal Cycling Programme

Step	*Temp*	*Time*	*Cycle*
Initial Denaturation	96°C	2 min	-
Denaturation	96°C	30 sec	Repeat 29x
Annealing	58°C*	30 sec	
Extension	67°C	1 min	
Final Extension	67°C	5 min	-
Rest	4°C	∞	-

Table 2.5: Tables Showing PCR reaction mix and Thermal Cycling programme for a standard PCR reaction. *Annealing temperature varies depending on the primers used, and was adjusted to 0.5-1.0°C below the lowest primer melting temperature.

2.5.3 Agarose Gel Electrophoresis

To check the size of dsDNA fragments generated e.g. by PCR or Restriction Endonuclease Digest, samples were run on an agarose gel.

Gels aimed to target fragments >1kb or <1kb were made with either 1% or 1.5% respectively of UltraPure™ Agarose (Invitrogen – 16500-500) in TAE Buffer (40mM Tris acetate, 1mM EDTA pH 8.0). The solution was heated for roughly 2 minutes in a microwave and allowed to cool before adding 0.2µg/ml Ethidium Bromide (Sigma – E1510) and pouring into a gel cast.

Gels were placed in a gel tank with TAE Buffer, and samples were diluted 5:6 with Purple Gel Loading Dye (NEB – B7025) before loading. Samples prepared using ReddyMix PCR Master Mix (as described in 2.5.2), already contained loading dye, and so no more was added. A DNA ladder of either 100bp (NEB – N3231) or 1kb (NEB – N3232) was also loaded as a marker of

fragment size. Gels were run at 100mV for 1 hour, or until sufficient size separation was observed. Gels were visualised using a UVP BioDoc-It™ Imaging System.

2.5.4 PCR Clean-Up

Prior to downstream processing of PCR products, it is frequently required to perform a 'clean-up', removing any remaining primers and dNTPs. For Sanger sequencing reactions, residual primers can cause bidirectional sequencing, resulting in overlapping sequence traces that are unreadable. The Sanger sequencing reaction mixture contains a specific ratio of dNTPs to ddNTPs, which would be altered by residual dNTPs from the PCR reaction.

In the case of TA cloning systems, primer dimer products can insert themselves into the plasmid, reducing cloning efficiency.

2.5.4.1 ExoSAP-IT

For Sanger sequencing of PCR products, 2.5µl of PCR product was added to 1µl ExoSAP-IT® PCR Product Cleanup (Affymetrix – 782011) and 11.5µl nuclease free H_2O to a final reaction volume of 15µl. The reaction mixture was run on an MJ Research PTC-200 thermal cycler at 37°C for 15 minutes, and then 94°C for 15 minutes to denature the enzymes.

2.5.4.2 Gel Extraction

For cloning of PCR products, gel extraction was used to clean up the PCR reaction. This allows size selection of the PCR product based on the section of the gel that is excised. Cloning efficiency improves as insert size decreases, and so smaller PCR products that may be present in quantities too low to observe on a gel, could still influence overall cloning efficiency, particularly when trying to insert large fragments.

For gel extraction, gels were made with Low Melting Point (LMP) Agarose (Promega – V3841), which is more efficient for DNA extraction, and the gel tanks were run on ice to prevent the gels from melting in the heat generated by the current.

Bands were excised from the gels using a scalpel, and DNA was extracted using a QIAquick PCR Purification Kit (Qiagen - 28706), following the recommended protocol.

2.5.5 Sanger Sequencing

Sanger sequencing was performed on PCR products after clean up, or on plasmid DNA after extraction. Due to the high copy numbers of plasmids after extraction, amplification by PCR is not required prior to sequencing. The sequencing reaction was performed using BigDye Terminator v2.1/v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific - 4337455). The reaction mix and the standard PCR programme are shown below in

Table 2.6. Each sample requires two reactions, one containing the forward primer, and one containing the reverse.

Sequencing Reaction Mix

Reagent	*Volume*
5X Sequencing Buffer	2µl
Sequencing Primer 20µM	0.4µl
BigDye Reaction Mix	0.5µl
Nuclease Free H_2O	4.6µl
Template DNA	2.5µl
Total	***10µl***

PCR Thermal Cycling Programme

Step	*Temp*	*Time*	*Cycle*
Initial Denaturation	96°C	1 min	
Denaturation	96°C	30 sec	Repeat 29x
Annealing	58°C*	15 sec	
Extension	62°C	1 min	
Rest	4°C	∞	

Table 2.6: Tables Showing Sanger sequencing reaction mix and Thermal Cycling programme for a standard sequencing reaction. *Annealing temperature varies depending on the primer used, and was adjusted to 0.5-1.0°C below the primer melting temperature.

After the sequencing reaction, DNA was purified by ethanol precipitation. 30µl of 100% ethanol and 100mM Sodium Acetate was added to each reaction, samples were then incubated at 4°C for 20 minutes before being centrifuged at 3060xg at 4°C for 20 minutes. The supernatant was tipped off, and 30µl 70% Ethanol was added. Samples were then incubated at 4°C for 5 minutes, before being centrifuged at 3060xg at 4°C for 10 minutes. The supernatant was tipped off again, and samples were left to air-dry at room temperature for 20 minutes.

Samples were then resuspended in 10µl Hi-Di™ Formamide (Thermo Fisher Scientific - 4404307) and incubated at 94°C for 2 minutes, before being analysed on a 3730xl DNA Analyzer.

Sequencing traces were analysed manually using SnapGene Viewer[387], and multiple sequence alignment analysis was performed using the online tool MUltiple Sequence Comparison by Log-Expectation (MUSCLE)[388,389].

2.5.6 Site Directed Mutagenesis (SDM)

SDM is an efficient technique for introducing small changes into plasmid DNA, and is particularly useful for regions with no restriction enzyme cleavage sites, or where there is no available template DNA for the desired sequence. It also allows greater flexibility when inserting new genetic features, and can even be used to insert new cleavage sites to ensure that subsequent cloning steps are kept in frame with the rest of the gene.

SDM was performed using the Q5® Site-Directed Mutagenesis Kit (NEB – E0554S). The reaction mix and the standard PCR programme are shown below in Table 2.7.

SDM PCR Reaction Mix

Reagent	*Volume*
2X Q5 Hot Start Master Mix	12.5µl
Forward Primer 10µM	1.25µl
Reverse Primer 10µM	1.25µl
Template Plasmid	10-20ng
Nuclease Free H_2O	Up to 10µl
Total	***25µl***

SDM PCR Thermal Cycling Programme

Step	*Temp*	*Time*	*Cycle*
Initial Denaturation	98°C	30 sec	
Denaturation	98°C	10 sec	Repeat 24x
Annealing	60°C*	30 sec	
Extension	72°C	30 sec/kb	
Final Extension	72°C	2 min	
Rest	4°C	∞	

Table 2.7: Tables Showing SDM PCR reaction mix and Thermal Cycling programme. *Annealing temperature varies depending on the primer used, and was adjusted to that recommended by NEBaseChanger[386] when the primers were designed. The extension time was calculated based on the size of the plasmid, with 30 seconds per 1000bp.

After amplification of the new plasmid sequence by PCR, a KLD reaction is used to circularise the product. 1µl of PCR product was added to 5µl of 2x KLD Buffer, with 3µl of nuclease free H_2O and 1µl of 10X KLD Enzyme Mix. The KLD reaction is incubated for 5 minutes at room temperature, before being transformed into competent cells for screening (as described in 2.5.8).

2.5.7 TA Cloning

Many of the polymerases used for PCR add leave a single 3'-A overhang, allowing easy ligation into plasmids with a 5'-T overhang. This is known as TA cloning, and is a useful tool to allow sequencing of individual DNA molecules from a pool of amplified fragments, this is useful for sequencing of genomic DNA where frameshift insertions or deletions make it difficult to determine the exact sequence, or for haplotyping of multiple SNPs observed within the same amplicon. It is also useful for locus specific bisulphite sequencing, where due to inaccuracies in the bisulphite conversion process, a percentage of methylation at each CpG is estimated based on the sequences of multiple bisulphite converted DNA molecules from the same sample.

TA cloning was performed using the pGEM®-T Easy Vector System I (Promega – A1360). pGEM®-T Easy Vector comes as a linearised plasmid with 5'-T overhangs, the plasmid contains an ampicillin resistance gene allowing for positive selection by ampicillin. The insertion site is within a *lacZ* operon that is disrupted if a fragment is successfully inserted, allowing for blue/white colony selection on agar plates containing IPTG & X-gal.

PCR products for ligation were purified by gel extraction (as described in 2.5.4.2), and then 3µl were added to 5µl of 2X Rapid Ligation Buffer, 1µl pGEM®-T Easy Vector and 1µl T4 DNA Ligase for a final ligation reaction volume of 10µl. The ligation reaction was incubated at 4°C overnight, and transformed into competent cells for screening (as described in 2.5.8).

2.5.8 Bacterial Transformation for Plasmid Expansion, Colony Separation & Glycerol Stocks

Plasmids were transformed into E. coli DH5α competent cells from stocks maintained by our laboratory group. 5 – 10µl of plasmid solution were added to 50µl DH5α, and incubated on ice for 30 minutes, before heat shock at 42°C for 45 seconds, and were then immediately returned to ice for 5 minutes. 100µl S.O.C media (Invitrogen 15544034) was added, and cells were shaken at 37°C for 1 hour.

For plasmid extraction by maxiprep, cultures were then added to 50ml of LB broth containing either 50µg/ml Kanamycin or 20µg/ml Ampicillin, depending on the antibiotic resistance gene contained within the plasmid. Cultures were then shaken at 37°C overnight.

To generate glycerol stocks for long term storage, the S.O.C cultures were added to 5ml LB broth containing either 50µg/ml Kanamycin or 20µg/ml Ampicillin, and shaken at 37°C overnight. For each stock, 500µl was mixed with 500µl of 50% glycerol (autoclaved), to a final concentration of 25% glycerol, and stored at -80°C.

For individual colony separation, S.O.C cultures were plated on LB agar plates, containing either 20µg/ml Ampicillin, 84µM IPTG & 40µg/ml X-Gal, or 50µg/ml Kanamycin. Plates were then incubated at 37°C overnight.

2.5.9 Colony PCR

Successful colonies were identified on agar plates after overnight culture, screened by resistance against the antibiotic. For the pGEM-T Easy plasmids, white/blue colony selection was used to screen for successful disruption of the *lacZ* gene (as described in 0).

Picked colonies were added to 50µl LB broth containing either 50µg/ml Kanamycin or 20µg/ml Ampicillin and incubated at 37°C for 3 hours. 1µl of the culture was then used as the DNA input for PCR (as described in 2.5.2), with an initial step of 96°C for 6 minutes.

The remaining culture was stored at 4°C until after screening by Sanger sequencing.

2.5.10 Plasmid Extraction from Bacterial Cultures

For plasmids to be used for transfections, large quantities of DNA are required. These plasmids were extracted from 50ml overnight cultures (as described in 2.5.8) using a QIAGEN Plasmid Maxi Kit (Qiagen - 12163), following the standard protocol.

For plasmids undergoing multiple cloning steps during plasmid construction, only a small amount of DNA is required from successful colonies to progress to the next stage. For these colonies screened by colony PCR, the 50µl stock at 4°C (as described in 2.5.9) was added to 10ml LB broth containing either 50µg/ml Kanamycin or 20µg/ml Ampicillin and shaken overnight at 37°C. Plasmids were then extracted using Wizard® *Plus* SV Minipreps DNA Purification System (Promega – A1330), using the centrifugation protocol.

Concentrations were analysed by NanoDrop Spectrophotometer (NanoDrop 2000 or NanoDrop One).

2.5.11 cDNA Conversion & Analysis

1µg of RNA was converted to cDNA using either SuperScript™ II Reverse Transcriptase (Thermo Fisher Scientific - 18064014) or ProtoScript® First Strand cDNA Synthesis Kit (NEB – E6300S), following the recommended protocols for using poly(T) primers. Each sample was also run with a negative control tube lacking the Reverse Transcriptase enzyme (RT negative).

For real-time PCR (rt-PCR) analysis, Power SYBR Green PCR Master Mix (Thermo Fisher Scientific - 4368702) was used, with a 50µl reaction mix containing 1µl sample cDNA, 0.5µl of each 20µM primer, 23µl nuclease free H_2O and 25µl SYBR Green Master Mix. Reactions were run with β-actin as an endogenous control, and both H_2O and RT negative samples as negative controls. Where possible, triplicate biological replicates were used, and each individual cDNA sample was run in duplicate as a technical replicate.

Reactions were set up on MicroAmp® Fast Optical 96-Well Reaction Plates (Thermo Fisher Scientific - 4346906), and were run on an ABI 7900HT Real Time PCR System. The primers used for rtPCR analysis are shown in Appendix 4.

2.6 Cell Culture Conditions

2.6.1 K562 Growth Conditions

K562 is a human erythroleukaemic cell line, derived from a female chronic myelogenous leukaemia patient in 1970[390]. K562 cells were generously provided by Professor Thein's laboratory.

K562 cells were grown at a concentration of between $2x10^5$ – $1x10^6$ cells per ml or medium, roughly doubling each day, and splitting every 2-3 days. Cells were grown in RPMI 1640 Medium (Thermo Fisher Scientific - 21875091) with 10% FBS (Thermo Fisher Scientific - 10270106) and 1% Penicillin/Streptomycin (Thermo Fisher Scientific - 15140122), at 37°C with 5% CO_2.

2.6.2 Freezing & Thawing

Aliquots of $1x10^6$ cells were frozen for storage in 1ml of freezing solution. Freezing solution consisted of 90% FBS (Thermo Fisher Scientific - 10270106) & 10% DMSO (Santa Cruz Biotechnology – sc-202581). Vials were then transferred to a Mr. Frosty™ Freezing Container (Thermo Fisher Scientific – 5100-0001) filled with Isopropanol, and incubated at -80°C for 1 – 3 days, before being transferred to liquid nitrogen for long term storage.

Frozen aliquots were thawed by warming rapidly in a 37°C water bath, and immediately diluting with 10ml pre-warmed growth media in a drop-wise manner. Cells were then centrifuged at 1800rpm for 5 minutes, and the medium fully replace before plating.

2.6.3 DNA & RNA Extractions

DNA was extracted from K562 cell culture using a Qiagen DNeasy Blood & Tissue Kit (Qiagen - 69504), ideally using $>2x10^6$ cells. In the case of cells that had been sorted by FACS, the number of cells was typically lower than this, and a Qiagen QiaAMP DNA Micro Kit (Qiagen - 56304) was used.

RNA was extracted using a Qiagen RNeasy Mini Kit (Qiagen - 74104), performing the optional DNase treatment step using RNase-Free DNase Set (Qiagen - 79254). Both DNA & RNA concentrations were assessed by either NanoDrop Spectrophotometer (NanoDrop 2000 or NanoDrop One) or by Qubit using Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific -

Q32854) and Qubit® RNA HS Assay Kit (Thermo Fisher Scientific - Q32852) for DNA and RNA respectively.

2.6.4 Transfections

Three different transfection techniques were tested, with varying success in terms of transfection efficiency.

2.6.4.1 Lipofectamine 2000

$2x10^6$ cells were seeded in a 6-well plate in 2ml growth medium (without antibiotics) and incubated overnight. Between 2-10µg plasmid DNA was mixed with 250µg RPMI 1640 (without antibiotics or FBS), in a separate tube 10µl Lipofectamine 2000 (Thermo Fisher Scientific - 11668027) was mixed with 250µl RPMI 1640 (also without antibiotics or FBS). Mixtures were incubated at room temperature for 5 minutes, and then combined to make a transfection solution of 500µl, which was incubated at room temperature for 20 minutes.

The transfection was then added to the cells in the 6-well plate, and then returned to the incubator. After 5 hours, cells were washed and replated in fresh growth medium, complete with both antibiotics and FBS.

2.6.4.2 Calcium Phosphate

Transfections were performed using the Calcium Phosphate Transfection Kit (Sigma-Aldrich - CAPHOS). $4x10^5$ cells were seeded in a 6-well plate in 2ml of complete growth medium and incubated overnight. A full media change was then performed two hours before transfection. 12µg plasmid DNA was added to 15µl 2.5M Calcium Chloride, and made up to 150µl with nuclease free water and mixed thoroughly by pipetting.

150µl 2X HEPES buffered saline pH 7.05 was added to a separate tube, and the 150µl Calcium Chloride:DNA solution was added in a drop-wise manner, whilst the HEPES solution was gently agitated by passing air through it with a 1ml pipette. This transfection solution was then incubated at room temperature for 20 minutes, before being added to the cells in the 6-well plate and gently mixed. Plates were then returned to the incubator for 16 hours, after which cells were washed and replated in fresh growth medium.

2.6.4.3 Nucleofection

Nucleofection was performed using Amaxa® Cell Line Nucleofector® Kit V (Lonza – VCA-1003), and was run on a Nucleofector™ 2b Device (Lonza – AAB-1001).

Cells were split 24 hours in advance of nucleofection, to ensure that they were in growth phase. $1x10^6$ cells were resuspended in 100µl Nucleofector® Solution, and 1-5µg plasmid DNA, 30µg ssODN or 30pmol siRNA were added either individually or in combination, in a maximum volume of 10µl (10% of Nucleofector® Solution). This reaction mix was immediately transferred to a cuvette, ensuring that the solution completely covered the gap between the metal plates, before placing in the nucleofector and running on programme T-016.

Upon completion of the programme, 500µl of complete growth medium was added to the cuvette, and the cell suspension was immediately transferred to a 6-well plate, containing another 2ml of medium.

2.6.5 Positive Selection for Plasmid Uptake & Clonal Expansion

48 hours after transfection, cultures were sorted based on GFP expression by FACS. FACS was carried out on BD FACSAria™ cell sorters, as a service provided by the BRC Flow Cytometry Core Facility at Guy's Hospital. Cells were sorted onto 96-well culture plates, with one cell per well in 200µl of complete culture medium, to allow clonal expansion of gene edited cell lines.

From the seventh day after FACS, cultures were checked every two days, looking for signs of colony growth. Identified cultures were transferred to 6-well culture plates and grown until they reached a concentration of approximately $1x10^6$ cells per ml, after which they were cultured under the standard K562 culture conditions described in 2.6.1, with DNA and RNA extracted as described in 2.6.3.

Cells were screened for the SNPs of interest by Sanger sequencing following PCR amplification using the primers shown in Appendix 5.

Chapter 3 Results: Erythroid Progenitor Isolation

3.1 In vitro Culturing of Erythroid Progenitors

The rationale for this part of the thesis work was to use an *in vitro* culture system to develop erythroid progenitor cells from the peripheral blood of SCA patients, to allow us to isolate a late stage progenitor population prior to enucleation. This would allow investigation into DNA methylation and other epigenetic marks with important roles in gene regulation in these cells, conducting longitudinal studies to investigate how these marks are affected by different treatments, such as HU therapy.

As discussed in 1.4.3, there are a variety of different in vitro culture techniques currently used for expansion and differentiation of erythroid cells from PBMCs *in vitro*. The technique outlined in 2.1.2 is based on a culture system that has been demonstrated to maximise the number of cells, and does not rely on the expensive pre-selection of $CD34^+$ cells[159]. This technique was routinely used by Professor Thein's laboratory group, and they had extensive experience using this culture system to grow cells from healthy donors.

3.1.1 Healthy Donor Blood Culturing

Initially, PBMCs were cultured from healthy blood donors. This approach was chosen because the erythroid precursors in the culture are particularly sensitive, and it was important to demonstrate that the culturing process was robust and reproducible before attempting to culture SCA blood samples collected from consenting patients in clinic.

This was especially important since much smaller volumes of blood were available from the SCA patients than the 50ml that was routinely used for this culturing process. Since SCA is a haemolytic disorder, the volume of blood collected from patients in the clinic has to be minimised, and only between 9 – 27 ml was available for culturing, as dictated by the study protocol. It was anticipated that a reduced starting cell number would compromise the viability of the culture.

During Phase 1 of the culture, the number of cells decreased rapidly as expected, with the majority of the cell populations that make up the PBMC layer undergoing cell death while the erythroid precursor population expanded. By P1D6, this resulted in a relatively pure cell culture with some monocyte contamination, 20% – 50% of the size of the initial PBMC sample (Figure 3.1).

After the switch to Phase 2 at P1D6, the majority of cultures did not recover, and the cell count remained in decline, although some signs of differentiation were visible. Figure 3.1 shows growth curves of four of these cultures, only one of which was successfully expanded after entering Phase 2, yielding more cells than were plated at P1D0. As expected, successful cultures differentiated while expanding, resulting in a population that was relatively homogenous in terms of both cell lineage and developmental stage (Figure 3.2 & Figure 3.3). Isolation of this population would allow for much more sensitive analyses than is possible when working on the mixture of PBMCs that were initially isolated from the blood sample.

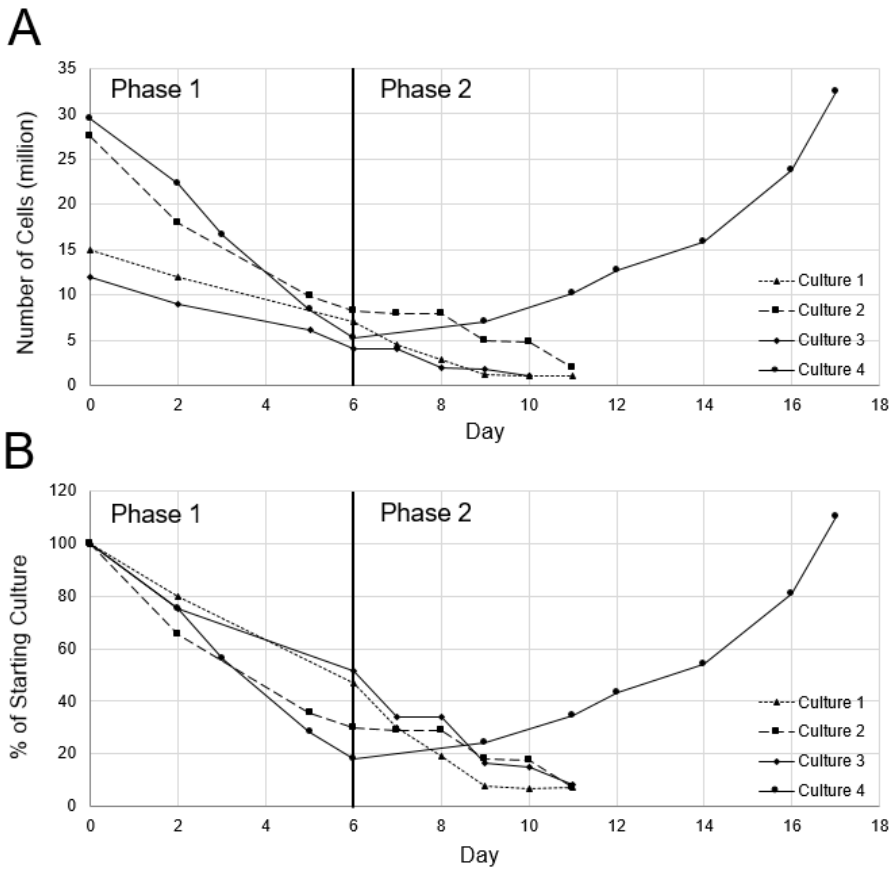


Figure 3.1: Growth curves showing the progress of erythroid cultures from healthy blood PBMCs. A – Growth as total number of cells. B – Growth as a percentage of the starting cell number at P1D0. The black line at Day 6 indicates the transition from Phase 1 to Phase 2, and can be considered as both P1D6 & P2D0. Of the four cultures, only Culture 4 successfully recovered and expanded after switching to phase 2. Cultures 1-3 continued to experience large amounts of cell death, until being terminated early with only 1-2 million cells remaining, less than 10% of the starting culture.

Contrary to our expectations, the four cultures shown in Figure 3.1 are representative of the poor success rate that was experienced when culturing these cells, which appeared to vary both between individuals, as well as over time. Interestingly, the rate of reduction in cell number as erythroid precursors expand and other blood populations are lost during phase 1, seems to bear no indication as to the likelihood of success of the culture. In Figure 3.1, the successful culture

actually experienced the greatest reduction, down to roughly 20% of the initial starting population.

It is worth noting that the cultures appeared to be sensitive to the age of the reagents used, and to SCF in particular. While this data was not recorded, cultures performed with freshly ordered SCF appeared to be healthier than those using aliquots a couple of months after delivery. This is despite the fact that these had been divided into small aliquots and stored at -20°C upon arrival, to minimise the number of freezing and thawing cycles that each aliquot experienced. It is not clear why this occurred, but it is thought that it may be the result of a faulty freezer, and that perhaps storage temperature was not as consistent as was expected. Upon noticing this effect, SCF aliquots were replaced more often, although the frequency at which new vials were ordered was limited by the fact that SCF is by far the most expensive reagent of the culture.

The progress of each culture was assayed by cytospin daily throughout Phase 2, as shown in Figure 3.3. When cells reached the polychromatic erythroblast stage of development (typically around P2D7 - P2D10), $CD71^{+}GPA^{+}$ cells were isolated by FACS (Figure 3.2) in order to guarantee the purity of the population in terms of developmental stage. FACS also ensured removal of any monocyte populations that frequently persisted to this stage, and were the main source of contamination by non-target cells.

Figure 3.2 shows flow cytometry data of a successful culture at P2D9 compared to freshly isolated PBMCs. Although some variation in cellular composition would be expected, since the samples are from two different individuals, it is clear that the cultured cells have been greatly enriched for $CD71^{+}$ erythroid progenitors, and that the $CD45^{+}$ leukocyte populations have been reduced. CD71 & GPA expression from these data can be used to assess the developmental stage, and while the majority of cells are still $CD71^{+}GPA^{-}$, some GPA expression is observed, and it would be expected that shortly after this the majority of cells would have entered the $CD71^{+}GPA^{+}$ stage, before starting to lose CD71 expression during the terminal stages of differentiation.

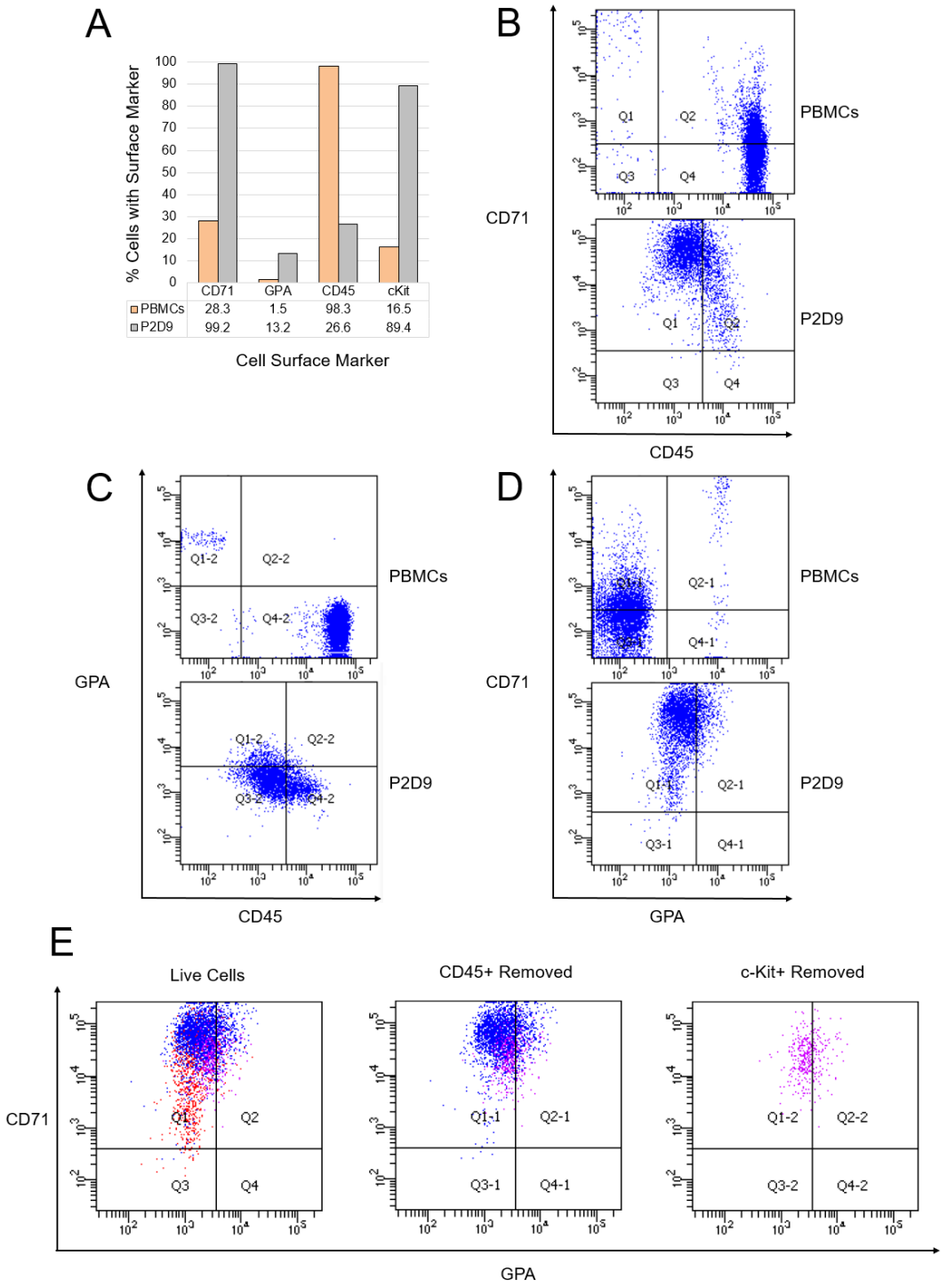


Figure 3.2: Flow Cytometry data from healthy PBMCs directly after isolation, compared to at P2D9 of a successful culture. Samples are from two separate healthy donors. A – Percentage of cells positive for each of the four cell surface markers: CD71, GPA, CD45 & cKit. CD71 & cKit are greatly enriched in the P2D9 cells compared to the PBMCs, increasing to 99.2% & 89.4% respectively. $CD45^+$ cells are reduced to 26.6% in the cultured sample, making up 98.3% of the PBMCs. B – CD71 & CD45 plots. CD45 & CD71 are co-expressed be some cell populations in both samples, although the majority of cells express either CD71 or CD45. C – GPA & CD45 plots. There is no overlap in expression of GPA & CD45 in either sample, as is expected given the specificity of GPA as a late stage erythroid marker. D – CD71 & GPA plots. Two distinct but faint GPA^+ populations are present in the PBMC sample; $CD71^+$ and $CD71^-$. Loss of CD71 expression marks the transition to a later stage of erythroid progenitor development. In the cultured sample, only the $CD71^+$ population is observed. E – Effect of FACS filtering gates on CD71 & GPA plot of P2D9 cultured cells. Red, blue & magenta represent $CD45^+$, c-Kit^+ and $CD45^-$c-Kit^- cells respectively. The position of the $CD45^-$c-Kit^- population shows that the culture is differentiating, as the $CD71^+$ cells start to express GPA.

Interestingly, c-Kit also seems to be enriched in the cultured cells. c-Kit is typically a marker of early stage erythroid development, with a key role HSC self-renewal and quiescence, and expression is lost during erythroid maturation[160,391–393]. It is likely that this is an artefact of the culture system, since persistent expression of c-Kit has previously been associated with stress erythropoiesis, and has been observed both *in vivo* and *in vitro*[159,394]. If the biological activity of the SCF in the culture was in fact impaired, as was mentioned previously, this could also partially explain the increase of c-Kit in the cell surface, since SCF is the ligand for the c-Kit receptor.

The negative filtering of CD45 & c-Kit expressing cells is shown in Figure 3.2, demonstrating that the selection of $CD45^{-}c\text{-}Kit^{-}CD71^{+}GPA^{+}$ cells by FACS allows isolation of a homogenous population, and provides data for accurate developmental staging of the progenitors isolated.

Figure 3.3 shows the successful culture of erythroid progenitor cells. Two waves of differentiation are observed, the first occurring shortly after transition into phase 2, where a population of basophilic erythroblasts appears and is subsequently lost. The second wave of differentiation comes from a population of pro-erythroblasts that is maintained during the development of the first wave. This second wave of differentiation progresses further through the developmental pathway, and is responsible for the increase in cell numbers observed during the latter stages of the culture. It is thought that the early wave of differentiation is triggered by the increased concentration of pro-erythroblasts accumulated towards the end of Phase 1, in the absence of later stage erythroblasts. Mechanisms to address any imbalance between early and late stage erythroblasts would be expected as a normal part of erythropoietic homeostasis *in vivo*.

While it was shown that nucleated erythroid progenitors could be isolated from successful *in vitro* culture of healthy blood, the success rate of these cultures survival after transition to phase 2 was very low.

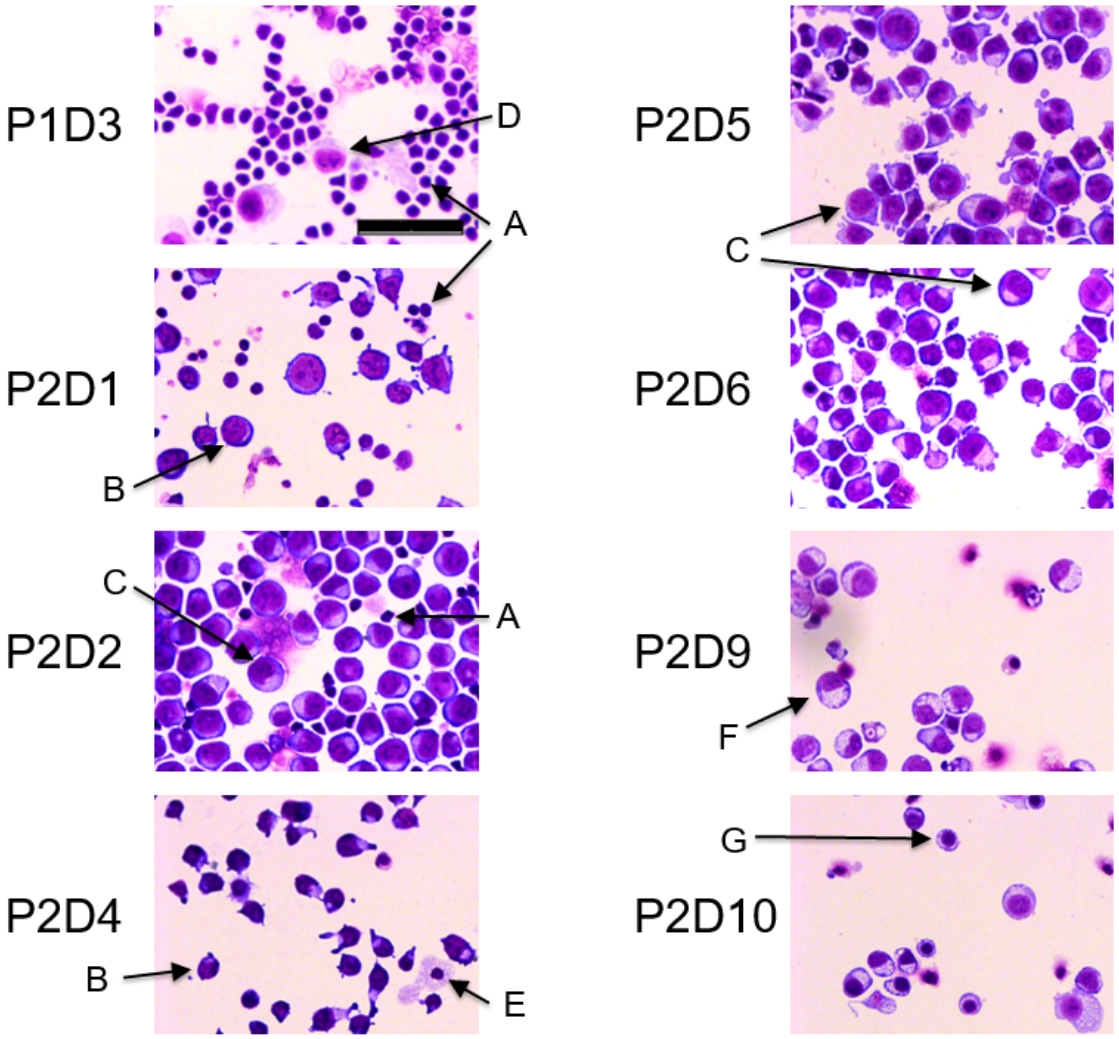


Figure 3.3: Photographs of cytospins showing *in vitro* culture of a healthy donor PBMC sample. Slides were stained with eosin & methylene blue. All photographs were taken at 40x magnification. The scale bar shown in P1D3 represents 50µm, and is the same for all photographs. A – Pro-erythroblasts, tightly packaged cells with no visible cytoplasm. B – Early basophilic erythroblasts, larger than pro-erythroblasts, cytoplasm can be seen to be expanding away from the nucleus. C – Late basophilic erythroblasts, much more of the cytoplasm is visible compared to early basophilic cells. D – White blood cell populations, distinguishable from erythroid progenitors by lack of staining around the cell membrane. E – Macrophage cell. F – Polychromatic erythroblasts, nucleus stains lighter, and cytoplasm appears larger, with more white space. G – Orthochromatic erythroblasts, nucleus is more condensed, and cytoplasm is smaller, as cells prepare for enucleation. An early wave of basophilic erythroblasts can be seen to appear at P2D2, and is lost by P2D4. Subsequently the proerythroblast population that persists at this stage starts differentiating and progresses through the erythroid developmental stages until the orthochromatic stage at P2D10.

3.1.2 SCA Patient Blood Culturing

The survival rate of erythroid progenitor cultures from healthy blood donors was unreliable, and sensitive to a variety of external factors. SCA patients experience increased stress erythropoiesis, with more early stage erythroid progenitor cells released into the peripheral blood. Therefore, it was thought that cells isolated from SCA blood samples might be more stable under culture conditions than blood from healthy donors. To test this, the technique was carried out on SCA (HbSS) patient peripheral blood.

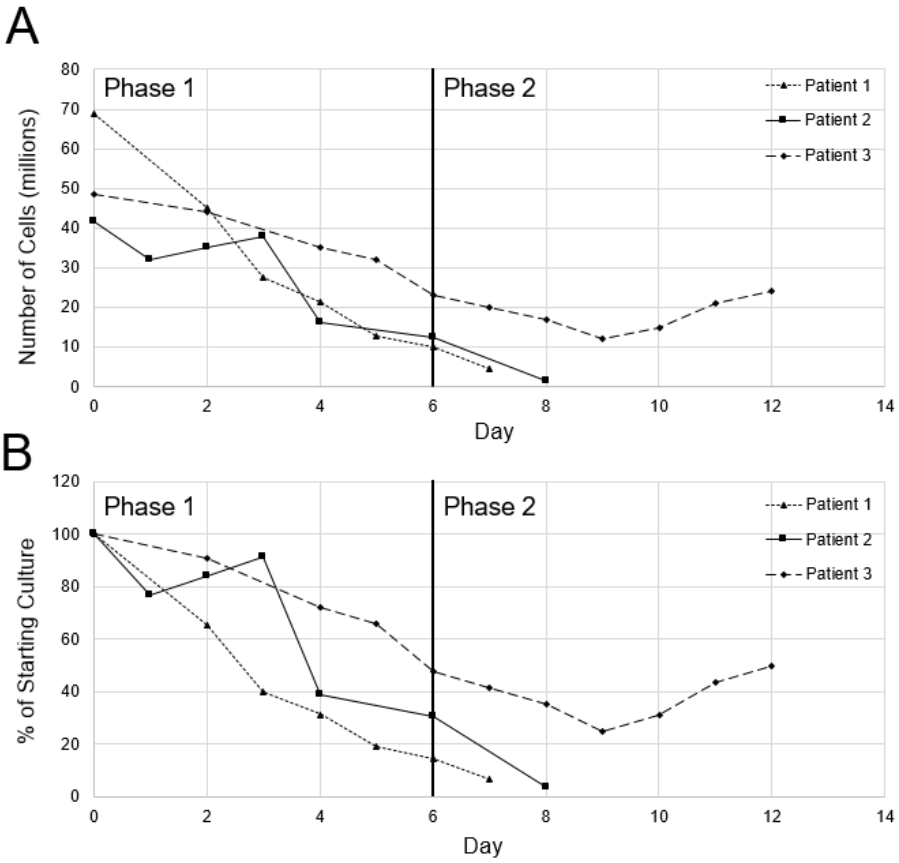


Figure 3.4: Growth curves showing the progress of erythroid cultures from SCA HbSS blood PBMCs. A – Growth as total number of cells. B – Growth as a percentage of the starting cell number at P1D0. The black line at Day 6 indicates the transition from Phase 1 to Phase 2, and can be considered as both P1D6 & P2D0. Only Patient Culture 3 successfully recovered after entering Phase 2, and this recovery was delayed, with growth not occurring until P2D4. Patient Culture 2 expanded early during Phase 1, dropping to 77% of the starting culture at P1D1, before steadily recovering to 91% at P1D3, and then dropping to 39% by P1D4. Note that Patient Culture 1 was divided and cultured as three separate sub-cultures, under the same conditions.

Survival rates of cells isolated from SCA (HbSS) patient blood were found to be just as unpredictable as healthy blood. Figure 3.4 shows growth curves for three patient blood sample cultures, and like the healthy blood samples shown in Figure 3.1, two out of the three cultures did not recover after entering Phase 2.

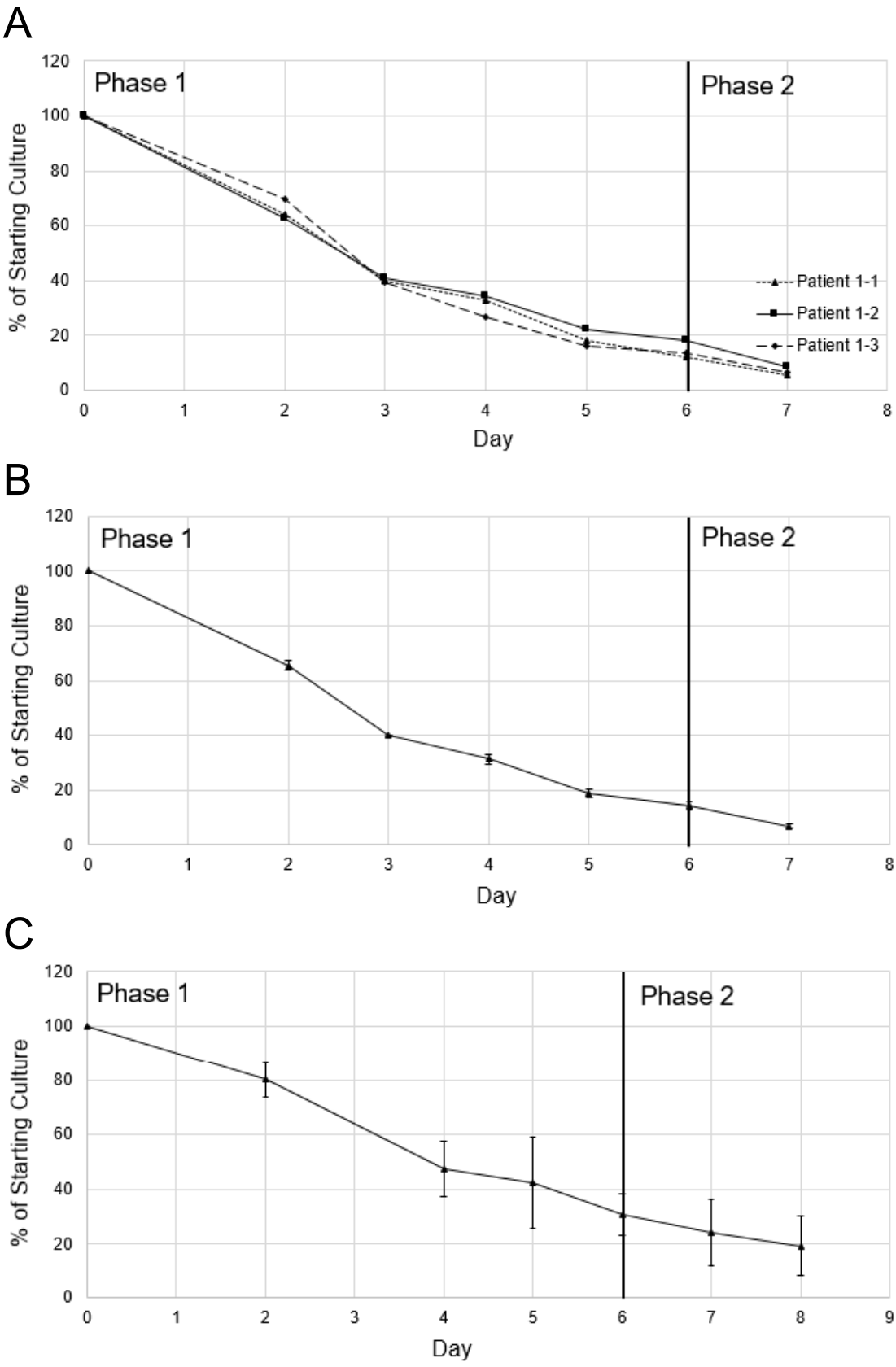


Figure 3.5: Growth curves showing the variability of erythroid cultures from SCA HbSS blood PBMCs. Patient Culture 1 from Figure 3.4 was divided into three sub-cultures at P1D0, and cultured concurrently in triplicate. A – Growth as a percentage of the starting cell number at P1D0. B – Mean of the growth curves shown in A, with error bars representing standard error. C – Mean of the growth curves shown in Figure 3.4. The black line at Day 6 indicates the transition from Phase 1 to Phase 2, and can be considered as both P1D6 & P2D0. The variation observed in the growth of the sub-cultures is very low, and much greater variation is observed between the cultures from different patients, cultured at different times.

In order to investigate the issue of reproducibility, the sample from Patient 1 in Figure 3.4, was divided into three sub-cultures at P1D0, and these were grown separately throughout the time course of the culture, the results of this are shown in Figure 3.5. The variation between the three cultures isolated from the same individual and grown concurrently show much less variation than is observed between the three separate HbSS patients cultured at different times. This difference in variation was to be expected during Phase 2, given that one of the samples eventually recovered, whilst the other two failed. More surprising was the difference in variation observed during Phase 1, with the triplicate cultures reducing in number at almost exactly the same rate. This suggests that rather than the non-erythroid lineages undergoing random cell death, the specific cellular composition of the PBMC sample and the cytokines produced by these cells determines a highly reproducible rate of decline.

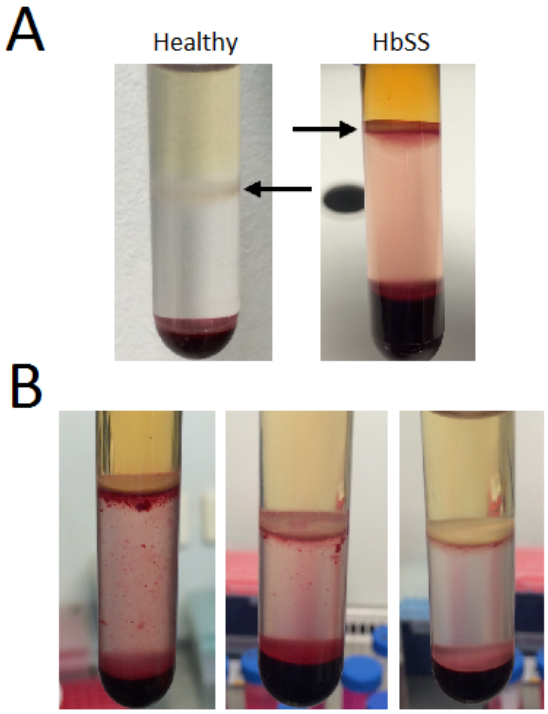


Figure 3.6: Photographs taken of PBMC layers, visible after density separation with Histopaque® - 1077. A – Comparison of HbSS & Healthy blood samples, arrows indicate PBMC layer. In HbSS patient blood samples, this layer appears red. B – Three additional HbSS samples. Variation in the thickness and the intensity of this red layer varies between patients.

Interestingly, the patient samples did behave differently under culture conditions compared to healthy donor blood. There is an additional population of cells that appears to be present in the HbSS blood samples, but not in healthy blood. These are believed to be a late stage erythroid population, present in the peripheral blood as the product of stress erythropoiesis. This population was quite unpredictable, and was maintained throughout phase 1, alongside the

erythroid precursors, while the other blood cells were lost. In Culture 2 of Figure 3.4 this nucleated red cell population appeared to expand rapidly during phase 1, and may be the cause of the growth in cell number observed at a stage when the culture was expected to reduce.

The presence of these cells was clearly visible after the separation of the Buffy Coat (Figure 3.6), with the PBMC layer appearing bright red, or with red layers rather than the white colour that is usually observed when processing healthy blood samples. The size and intensity of this red layer appears to vary between patients, and if these cells are a product of stress erythropoiesis occurring in the peripheral blood, then it would be expected that the presence of this population would be influenced by the severity of the SCA phenotype in each individual. This variation is observed in Figure 3.7, where blood from the milder HBSC genotype appears indistinguishable from a healthy blood sample.

Figure 3.7 shows the comparison between two patient samples, an HbSS genotype patient and a patient with the HbSC genotype, resulting in a milder form of the disease (discussed in 1.2.3.2). The flow cytometry analysis of the PBMC layer in the HbSS patient shows the presence of a large cell population that is available directly from peripheral blood and does not require *in vitro* culturing. Since this population is only present in the HbSS patient sample, it is thought to be the cell population that is visible as the red layer in HbSS PBMCs.

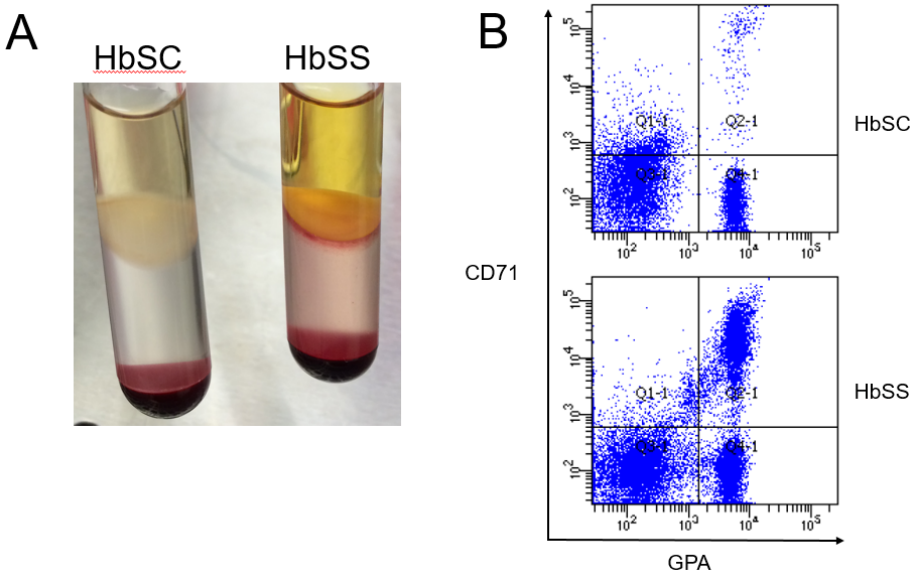


Figure 3.7: Comparison of PBMCs from an HbSC patient and an HbSS patient. A – Photograph of PBMC layers after density separation. The PBMC layer from the less severe HbSC patient does not have the red layer that is observed in HbSS patients, and is indistinguishable from a healthy PBMC layer (Figure 3.6). B – Flow Cytometry plots showing CD71 & GPA expression of the PBMC samples shown in A. The $CD71^{+}GPA^{+}$ cell population is present in both samples, but is more abundant in the HbSS PBMCs, making up 25.0% of cells, as opposed to 1.2% in HbSC. Both samples also have a high proportion of later stage CD71-GPA+ cells, 24.1% and 20.0% for HbSS & HbSC respectively.

3.2 FACS Isolation of Progenitors Directly from PBMCs

Due to the unreliability of the *in vitro* culture method for erythroid progenitor expansion, and the discovery of an erythroid progenitor population present in HbSS peripheral blood, sorting directly from PBMCs by FACS was tested.

3.2.1 FACS of Patient Blood Samples

Due to the logistics of collecting the sample from the clinic, and the FACS service only being available during working hours, PBMCs were not sorted directly after isolation from peripheral blood, but were kept in culture overnight, and sorted the following day at P1D1. It was thought that with such a short exposure to culture conditions, any influence on DNA methylation or transcription would be minimal, and any cell death induced in the immune cells would be beneficial, increasing the efficiency of the FACS process.

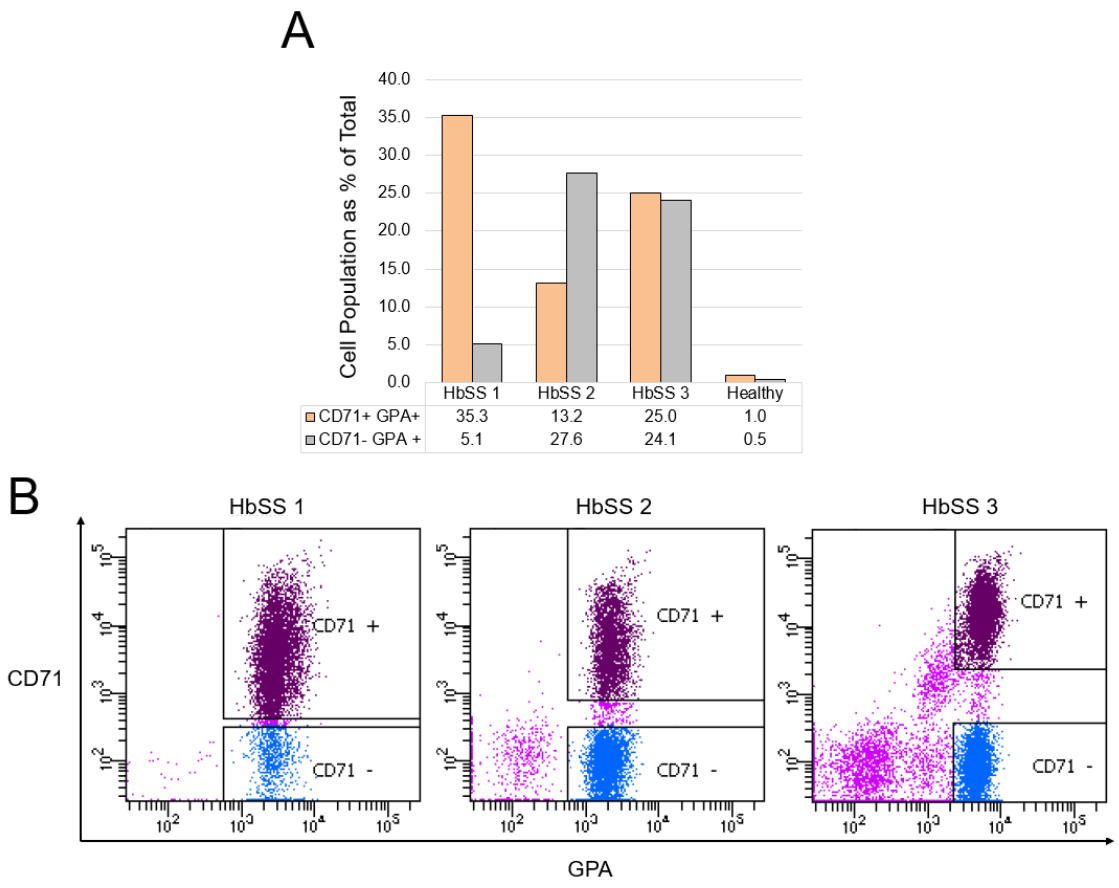


Figure 3.8: Flow cytometry analysis of three HbSS PBMC samples after <24 hours in culture. A – Numbers of $CD71^{+}GPA^{+}$ & $CD71^{-}GPA^{+}$ cells as a percentage of total PBMC layer, compared to a healthy PBMC sample. Levels of both populations vary between SCA patients, but are much higher than in the healthy blood sample. B – Flow cytometry plots of CD71 and GPA, after removal of CD45 and c-Kit, demonstrating the FACS gating used to collect each cell population. Magenta, maroon and blue represent $CD45^{-}CD14^{-}$, $CD71^{+}GPA^{+}$ & $CD71^{-}GPA^{+}$ cells respectively.

Figure 3.8 shows the flow cytometry analysis of cells collected from three HbSS patients. While the proportion of $CD71^{+}GPA^{+}$ & $CD71^{-}GPA^{+}$ cells varies between the HbSS samples, it is again clear that the $CD71^{+}GPA^{+}$ populations are specific to HbSS samples. Sufficient numbers of cells were acquired from the FACS process, with the $CD71^{+}GPA^{+}$ output reliably reaching more than $1x10^{6}$ cells. Given that only 500ng of DNA is required for the Infinium® HumanMethylation450 BeadChip to assay for genome-wide DNA methylation[395], and between 0.1-4.0µg of RNA is recommended for TruSeq Stranded mRNA Library Prep Kit for RNA-seq[396], these cell numbers were much higher than required.

3.2.2 DNA & RNA Extractions

Initially, it was intended that $CD71^{+}GPA^{+}$ cells would be stored in TRIzol Reagent at -80°C after isolation by FACS, with the aim of performing all DNA and RNA extractions in parallel after all the samples had been collected, so as to minimise variation. DNA & RNA extraction was tested for Sample 1 in Table 3.1, after storage in TRIzol for two months, and negligible amounts of both were obtained. Given that $6.5x10^{6}$ cells were originally collected, this was unexpected. For samples 2 & 3, cells were instead extracted immediately after sorting, using the Qiagen AllPrep Kit. Samples isolated immediately after extraction yielded measurable amounts of DNA & RNA, despite having less than half the input of Sample 1.

	Sample 1	*Sample 2*	*Sample 3*
Cell number	6.5×10^{6}	3.1×10^{6}	2.8×10^{6}
Extraction Method	TRIzol	Q-All	Q-All
RNA Concentration (ng/µl)	**<5.0**	11.2	13.0
DNA Concentration (ng/µl)	**<0.2**	4.5	**<0.2**
RNA Total Yield (µg)	**<0.15**	0.34	0.39
DNA Total Yield (µg)	**<0.02**	0.45	**<0.02**

Table 3.1: Three HbSS PBMC samples sorted on P1D1 by FACS. Q-All – Qiagen AllPrep DNA/RNA/Protein Mini Kit. Table shows the number of sorted cells, the method used to extract DNA & RNA, and the concentrations as assayed by Qubit. Sample 1 stored in TRIzol yielded negligible amounts of DNA & RNA, despite having the highest input cell number. DNA was also very low in samples 2 & 3.

While the extracted RNA is within the recommended range for RNAseq, the DNA fell below the 0.5µg recommended for the DNA Methylation array. While Sample 2 was close to this boundary, and may have still been successfully assayed, Sample 3 had negligible amounts of DNA. Most concerning was the fact that both the DNA & RNA were far below the yields expected. 1×10^{6} cells from cell lines typically yield >5µg and >10µg for DNA & RNA respectively[397], and although cell lines have different properties, and extractions typically have

higher yields than from primary tissue samples, this does not account for roughly 30 fold and 80 fold disparities for DNA & RNA concentrations respectively.

During the sample processing, after centrifugation of the samples post-FACS, it was observed that the cell pellets appeared smaller than expected for the given cell number. This, in combination with the low DNA & RNA yields, led to the conclusion that the majority of cells were not surviving the cell sorting process. FACS requires forcing cells through capillaries at high pressure, and dropping them into a collection tube one cell at a time. Both of these stages can lead to rupturing of the cell, and it is possible that the erythroid progenitors are too sensitive for this process to be used efficiently, even when run at the lowest pressure available on the FACS machine[398].

While it has been demonstrated in the literature that FACS can be used to isolate erythroid progenitors, for the most part this has been demonstrated in bone marrow or cord blood samples, which have a higher concentration of these progenitor cells[125,131]. A study that successfully isolated erythroid progenitors by FACS targeted cells at earlier stages (BFU & CFU, which are both $CD45^+$) and enriched the samples for $CD45^+$ cells using magnetic beads prior to sorting. They also observed that the cells isolated from peripheral blood had reduced ability to form colonies compared to those from cord blood[174].

3.3 Miltenyi BeadKit Isolation of CD71+GPA+ Progenitors from PBMCs

Due to the harsh conditions and high rate of cell death associated with FACS, isolation by Miltenyi BeadKit was tested. This technique relies on antibody-magnetic bead conjugates, and while it does not have the same specificity as FACS, where the sorting process occurs after analysis of each cell and so 100% purity is expected, it is a more gentle isolation process. Isolation of erythroid progenitors using the magnetic bead separation technique had previously been demonstrated by Walker *et al.*[245].

3.3.1 Enrichment for CD71+ Cells

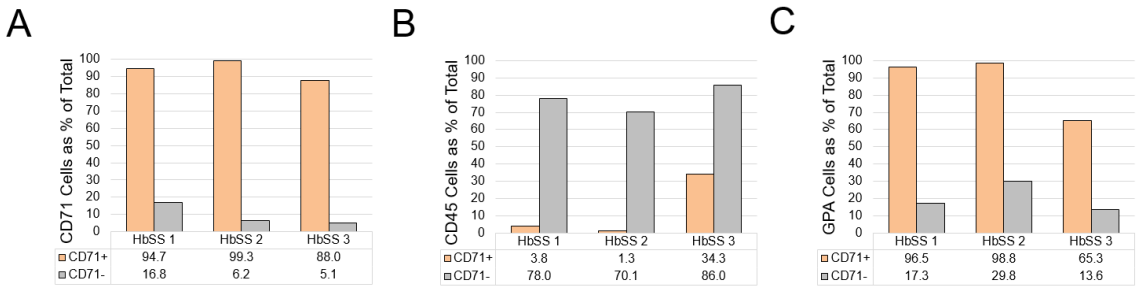


Figure 3.9: Flow Cytometry data from CD71 BeadKit enrichment of three HbSS patient PBMCs. Both the $CD71^+$ fraction (orange) and the $CD71^-$ fraction (grey) were analysed. A – CD71 staining. CD71 is successfully enriched in the $CD71^+$ fraction with a purity of 88.0 – 99.3%. B – CD45 staining. The $CD45^+$ cells that make up the majority of PBMCs are successfully reduced in the $CD71^+$ fraction, to <4% in HbSS 1 & 2, but only to 34.3% in HbSS 3. C – GPA staining. Similarly to CD71, GPA is successfully enriched in the $CD71^+$ fraction, to >96% in HbSS 1 & 2, but only 65.3% in HbSS 3.

Figure 3.9 shows successful enrichment for $CD71^+$ cells by Miltenyi BeadKit for three HbSS patient samples. The purity of the CD71+ fraction was high for all samples, varying between 88 - 99%. The GPA^+ & $CD45^+$ cells were enriched and depleted respectively, and interestingly these cell populations appear to be complementary to each other in the CD71+ fraction. This suggests that two distinct erythroid progenitor populations are being isolated, the $CD71^+GPA^+$, as well as the $CD45^+CD71^+$. The $CD45^+CD71^+$ population represents an earlier stage of erythroid development, since CD45 is expressed on HSCs and is lost shortly after erythroid lineage determination[399], which was confirmed by the absence of co-expression with GPA, since CD45 is lost from the cell surface significantly in advance of GPA expression, although co-expression has been observed under some in vitro culture conditions[169].

While this $CD45^+CD71^+$ population appears to be low in HbSS 1 & 2 (<4%) it was much higher in HbSS 3 (34.3%), significantly reducing the homogeneity of the sample. These early stage progenitors had been observed previously, both in the cultured cells and PBMCs (Figure 3.2). In

these cases, the presence of the $CD45^+CD71^+$ cells was not a concern, since the FACS selection process included negative selection for CD45. It was therefore decided to use a CD45 Miltenyi BeadKit as a depletion step prior to enrichment for CD71.

3.3.2 CD45 Depletion Prior to Enrichment for CD71+ Cells

Figure 3.10 shows successful depletion of $CD45^+$ cells from the CD71+ fraction, and suggests that a reasonably reliable process for isolating the $CD71^+GPA^+$ cell fraction has been achieved. These results also stress the importance of testing purity of the isolated cell population from each sample by flow cytometry, since the failed enrichment of sample HbSS 7 would otherwise have gone unnoticed. Errors in the isolation process result in a completely different cellular composition of the enriched sample, and if these samples are not identified and removed from the study, they would identify cell-type specific differences in DNA methylation and transcription, as opposed to differences caused by drug treatment.

$CD71^+$ cells were more abundant in the CD45+ fraction than the CD71- fraction, this is likely due to the $CD45^+CD71^+$ cells that were observed in the CD71+ fractions in Figure 3.9, that are now being successfully removed by the CD45 depletion step.

Sample HbSS 9 in Figure 3.10 was from a SCA patient being treated with HU. There were concerns that if these erythroid progenitors were present in the peripheral blood as a result of the clinical severity of SCA, then it might be expected that this population would disappear in patients undergoing treatment. There is no apparent variation in cell surface markers, and 3.7 million $CD71^+GPA^+$ cells were collected (Table 3.2), more than in most of the untreated samples. Although this is only one sample, an increase rather than a decrease in $CD71^+GPA^+$ cells may be expected in patients undergoing HU therapy, since HU has been linked to increased stress erythropoiesis[44,400].

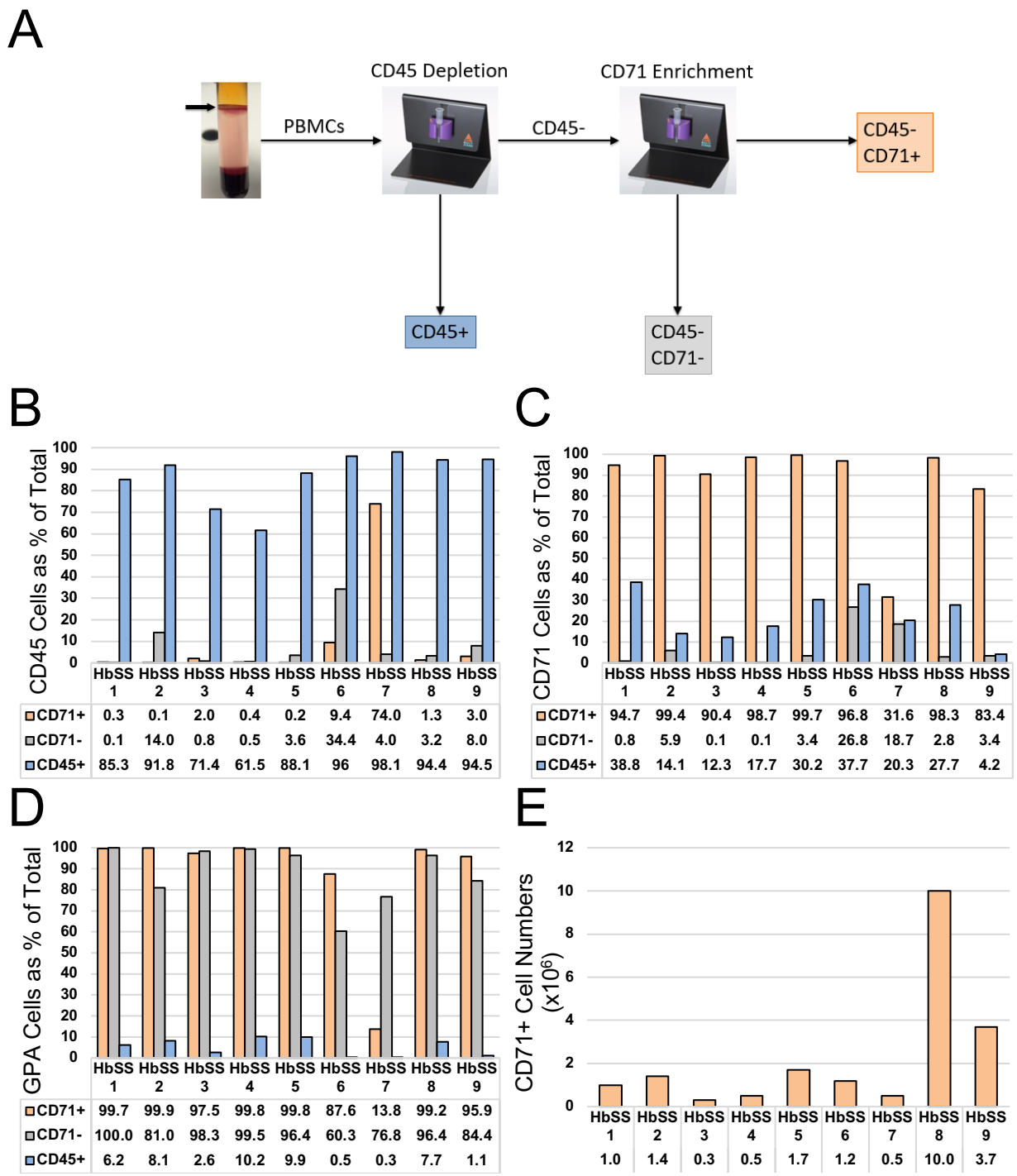


Figure 3.10: Flow cytometry data from CD71 enrichment of nine HbSS patient PBMCs, following CD45 depletion. A – Flow diagram illustrating the process of isolating the different cell fractions. The CD71+ fraction (orange), the CD71- fraction (grey) and the CD45+ fraction (blue) were analysed. Sample HbSS 9 was from a patient undergoing HU therapy. Percentage of cells stained in each fraction is shown for B – CD45, C – CD71 and D – GPA. Processing of sample HbSS 7 appears to have failed, with the CD71+ fraction containing only 31.6% $CD71^+$, 74.0% $CD45^+$ & 13.8% GPA^+ cells. Apart from HbSS 7, significant CD71 enrichment is observed in the CD71+ fraction for all samples, to between 80 – 99% purity. CD45 staining shows very low levels of $CD45^+$ cells in the CD71+ fraction, of between 0.1 – 9.4% (excluding HbSS 7). GPA staining confirms that the cells isolated in the CD71+ and CD71- fractions are the $CD71^+GPA^+$ and $CD71^-GPA^+$ cell populations respectively. E – Total cell counts of the CD71+ fraction from each sample, as estimated by haemocytometer counting. The total number of $CD71^+GPA^+$ cells isolated varied significantly between samples.

3.3.3 DNA & RNA Extractions

	HbSS 1	*HbSS 2*	*HbSS 3*	*HbSS 4*	*HbSS 5*	*HbSS 6*	*HbSS 7*	*HbSS 8*	*HbSS 9*
Cell number	1.0 x 10^6	1.4 x 10^6	0.3 x 10^6	0.5 x 10^6	1.7 x 10^6	1.2 x 10^6	0.5 x 10^6	10.0 x 10^6	3.7 x 10^6
Extraction	Q-All	Q-All	Q-All	Q-All	Q-All	Q-All	Q-All	Q-All	Q-Pure
RNA (ng/µl)	**<5.0**	166.0	11.1	45.6	92.4	17.4	**<5.0**	**>2000.0**	**DNA Only**
DNA (ng/µl)	**<0.2**	0.2	**<0.2**	**<0.2**	**<0.2**	0.9	2.3	0.8	**<0.2**
RNA Yield (µg)	**<0.15**	4.98	0.33	1.37	2.77	0.52	**<0.15**	**>60.0**	**DNA Only**
DNA Yield (µg)	**<0.02**	0.02	**<0.02**	**<0.02**	**<0.02**	0.09	0.23	0.08	**<0.02**

Table 3.2: Summary DNA & RNA extractions of the nine HbSS PBMC samples from Figure 3.10. Q-All – Qiagen AllPrep DNA/RNA/Protein Mini Kit. Q-Pure – Qiagen Puregene Blood Core Kit A. Table shows the number of sorted cells, the method used to extract DNA & RNA, and the concentrations as assayed by Qubit. <5.0 & <0.2 represent the lower limits of Qubit detection for RNA & DNA respectively, while >2000.0 represents the upper limit of RNA detection. RNA extraction was generally successful, and for sample HbSS 8, yielded more than was measureable by Qubit. Both DNA & RNA extraction failed for sample 1, and RNA extraction failed for HbSS 7. DNA extraction was unsuccessful, even in sample HbSS 8, with an input of 10.0 x 10^6 cells, which yielded >60.0µg of RNA. For HbSS 9, an alternative DNA extraction technique was tested with the entire sample of isolated cells, and was also unsuccessful.

Table 3.2 shows the total DNA and RNA isolated from the samples shown in Figure 3.10. RNA yields from the samples isolated by Miltenyi BeadKit was improved compared to the samples isolated using FACS, and larger cell pellets were observed after isolation, despite having a lower cell number. This suggests that less cell death is occurring during the isolation process, as is expected when avoiding the harsh conditions of FACS[398].

However, the DNA yield was very low, even in the sample that yielded >60µg RNA. The highest DNA output from the nine samples was only 230ng, roughly half of the recommended input for the DNA methylation analysis[395]. This came from sample HbSS 7, which had a high contamination of $CD45^+$ cells (Figure 3.10).

For HbSS 9, an alternative DNA extraction kit was tested (Qiagen Puregene Blood Core Kit A), this was used in Walker *et al.* for erythroid progenitor isolation, where they obtained sufficient DNA for locus specific bisulphite-sequencing to assess DNA methylation state. This was also unsuccessful, despite using the entire isolated sample as input for the DNA extraction, with no RNA extraction being performed.

3.3.4 Cytology: $CD71^+GPA^+$ Cells in HbSS Patients Are Enucleated

Due to the extremely low DNA yield from these cells, cytospins were prepared of the three fractions of an HbSS sample isolated by Miltenyi BeadKit, to allow visual identification of the stage of erythroid development, and to confirm that the cells were nucleated. These cytospins are shown in Figure 3.11, and demonstrate that the $CD71^+$ cells being isolated are at a later

stage of development than anticipated, and mostly consist of enucleated reticulocytes. This explains the imbalance in yield between the RNA & DNA extractions performed on these fractions.

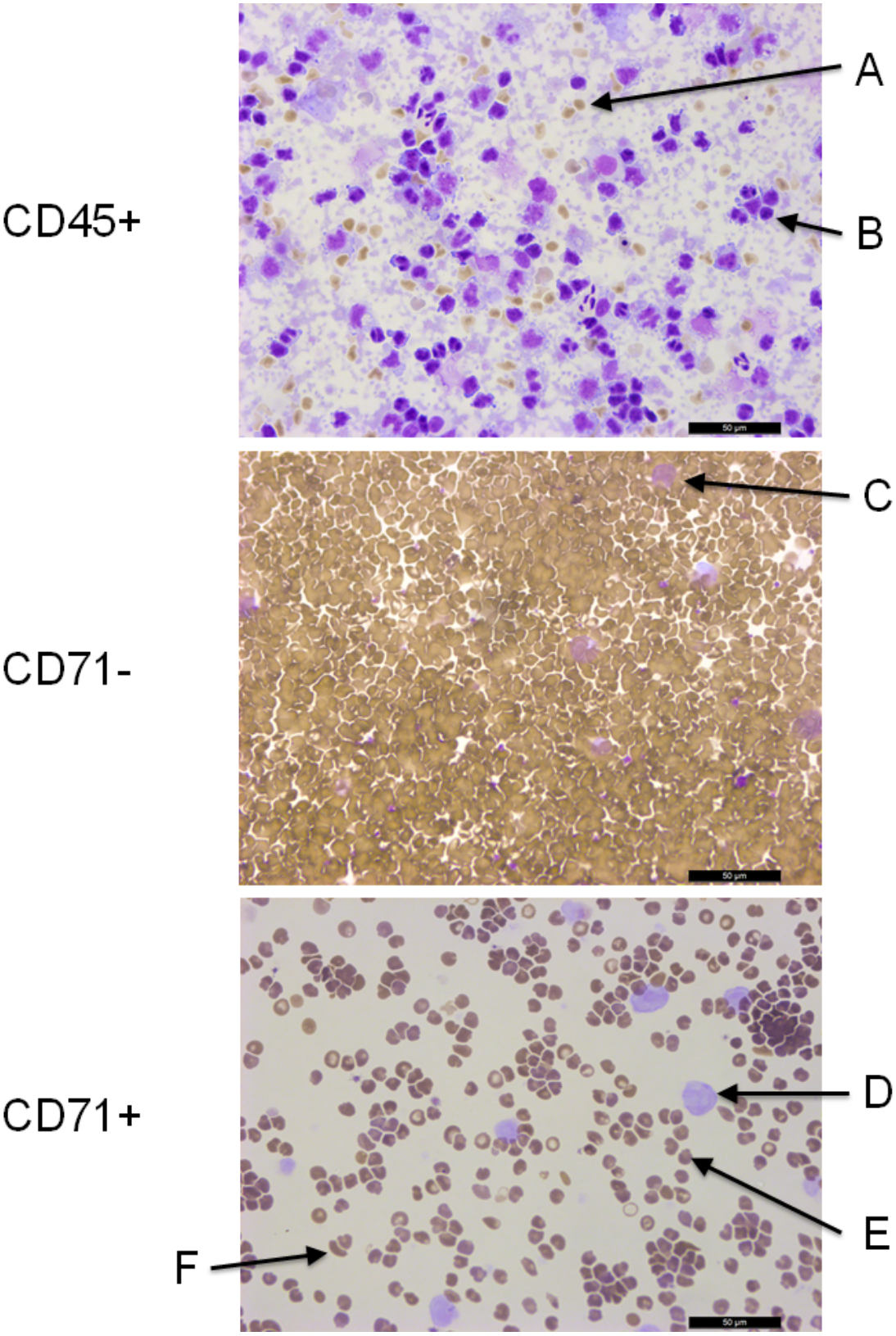


Figure 3.11: Photographs of cytospins taken from the three fractions of an HbSS patient blood sample isolated by Miltenyi BeadKit (CD45+, CD71- & CD71+). Slides were stained with eosin & methylene blue. Photographs were taken at 40x magnification, and scale bars represent 50µm. A – Red blood cell contamination in the CD45+ fraction. B – Nucleated CD45+ cells, nucleus stains as dark purple. C & D – Light purple staining indicates cytoplasm, but these cells are lacking a nucleus. E – Enucleated red cells. F – Sickling red cells. The CD45+ fraction mostly consists of nucleated PBMCs, with some red cell contamination. The CD71- fraction is densely packed with erythrocytes. The CD71+ fraction consists mostly of enucleated reticulocytes, staining slightly darker than in the other fractions. CD71+ & CD71- also contain larger enucleated cells, possibly post enucleation but prior to the reduction in volume that accompanies reticulocyte maturation[401].

3.4 Miltenyi BeadKit Isolation of Early Stage Progenitors from PBMCs

It is clear that the $CD71^{+}GPA^{+}$ cells isolated from peripheral blood have advanced to a later developmental stage than was anticipated, and in order to investigate epigenetic markers in the erythroid lineage, cells must be isolated at an earlier stage. It was therefore decided to attempt to isolate $CD34^{+}$ cells from peripheral blood, as has been successfully shown in the literature, and is routinely used to provide a pure erythroid precursor population for some of the *in vitro* culturing techniques[150,157,158,175]. $CD34^{+}$ progenitors are at an earlier stage of haematopoietic development, and would almost certainly be nucleated[399].

3.4.1 Enrichment for $CD34^{+}$ Cells

$CD34^{+}$ cells were isolated from peripheral blood by Miltenyi BeadKit, the same technique used to isolate $CD45^{-}CD71^{+}$ cells. Figure 3.12 shows $CD34^{+}$ cells from an HbSS patient that were successfully enriched to 97% in the CD34+ fraction. Similar to the CD71+ fractions isolated previously (3.3.1), there appear to be two distinct $CD34^{+}$ populations present, expressing either CD45 or GPA. As was observed previously, there was very little co-expression of GPA and CD45, which is expected given that CD45 expression is lost after the HSC stage in the erythroid lineage, and that GPA is a late stage erythroid developmental marker[399,402]. Interestingly, Figure 3.12C shows that $CD34^{+}$ isolation appears to enrich for the $CD34^{+}CD45^{+}$ subpopulation rather than $CD34^{+}GPA^{+}$, possibly due to higher cell surface expression of CD34 on these cells, inferred by the difference in CD34 staining intensity.

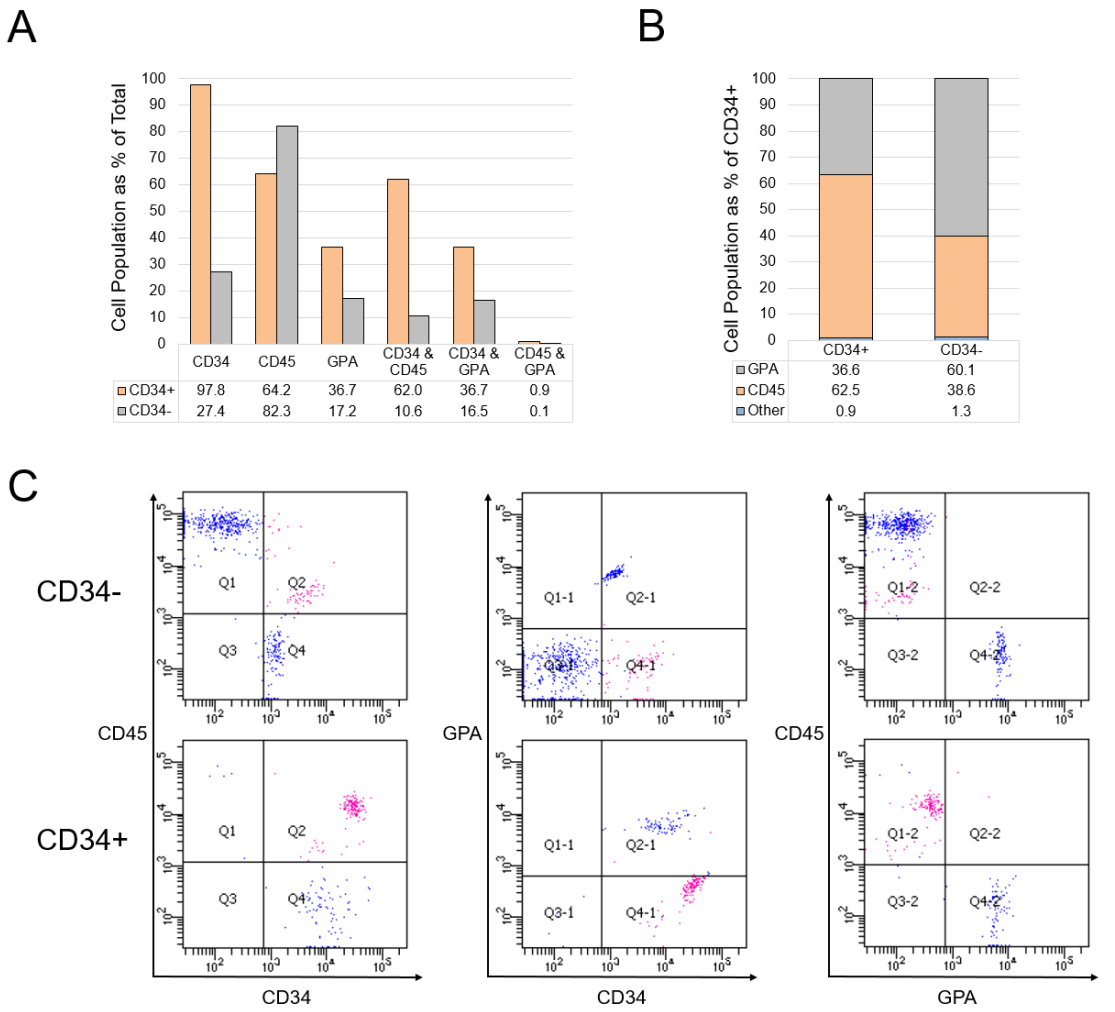


Figure 3.12: Flow Cytometry data from both fractions of an HbSS patient blood sample as isolated by BeadKit (CD34- & CD34+). A – Percentage of cells positive for each of the three cell surface markers: CD34, CD45 & GPA, as well as co-expression of each pair. $CD34^+$ enrichment was successful with 97.8% purity in the CD34+ fraction, compared to 27.4% in the CD34- fraction. B – Composition of $CD34^+$ population from both fractions. C – Graphs showing co-expression of the cell surface markers. Pink indicates $CD34^+CD45^+$ cells, as defined by gate Q2. Results indicate two distinct cell populations within the $CD34^+$ cells, with roughly 99% expressing either GPA or CD45, but <1% expressing both.

3.4.2 Cytology

Figure 3.13 shows cytospins of the CD34+ and CD34- fractions of an HbSS sample. Successful isolation of nucleated $CD34^+$ cells can be seen, and it is anticipated that these are the $CD34^+CD45^+$ cells and that the $CD34^+GPA^+$ cells are enucleated.

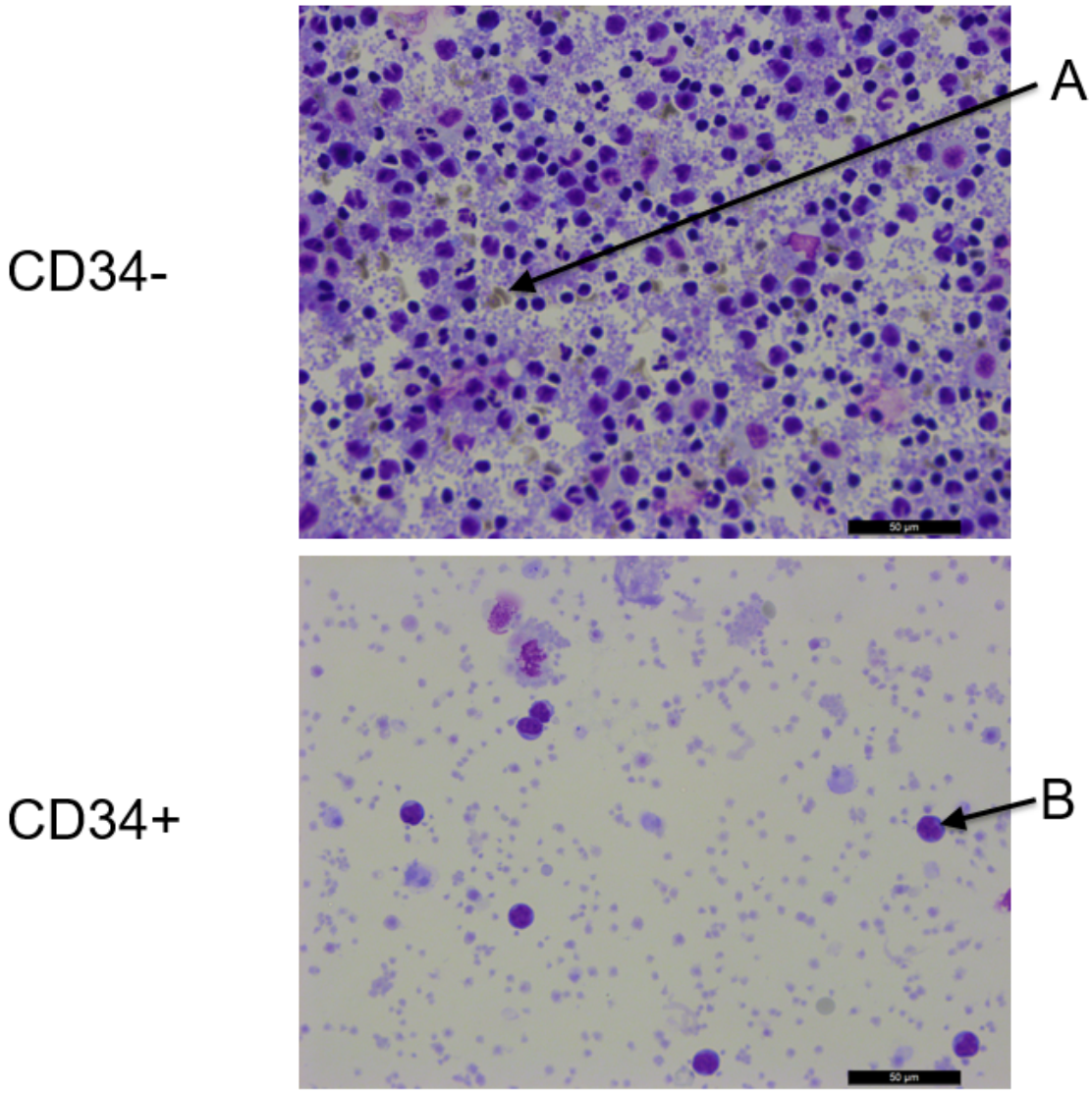


Figure 3.13: Photographs of cytospins taken from both fractions of an HbSS patient blood sample as isolated by BeadKit (CD34- & CD34+). Slides were stained with eosin & methylene blue. Photographs were taken at 40x magnification, and scale bars represent 50µm. A – Red blood cell contamination in the CD34- fraction. B – Nucleated $CD34^+$ cells. As expected the CD34- fraction contains the majority of the PBMC sample. The CD34+ fraction is less densely packed, and contains some debris and dead cells, as well as some cells lacking a nucleus. Nucleated $CD34^+$ cells are also visible.

3.4.3 DNA & RNA Extractions

Very low cell numbers were isolated in the CD34+ fraction of these samples, fewer than could be counted accurately by haemocytometer. Even when resuspended in 500µl of PBS, only one or two cells were visible in the counting chamber, suggesting that the cell number was in the order of $1x10^4$.

In order to obtain the results shown in Figure 3.12 and Figure 3.13, an entire sample was used for either flow cytometry or cytospin respectively. Even when a whole sample was used for DNA or RNA extraction, these extractions failed, falling below the threshold able to be detected by Qubit.

3.5 Miltenyi BeadKit Isolation of $GPA^{-}CD71^{+}$ Erythroid Progenitors

Since the number of CD34+ cells in circulation was found to be prohibitively low, another early stage population was investigated. Since the $GPA^{+}CD71^{+}$ cells were found to be enucleated, it was decided to deplete GPA^{+} cells, prior to a $CD71^{+}$ selection, with the aim of isolating GPA^{-} $CD71^{+}$ erythroid progenitors, which while less abundant would be expected to be nucleated. The flow cytometry analyses of the populations isolated by GPA depletion and CD71 enrichment are shown in Figure 3.14.

GPA^{+} cells were successfully removed from the sample, making up 89.0% of the GPA+ fraction, but only 0.4% and 1.7% of the GPA-CD71- and GPA-CD71+ fractions respectively. As had been observed previously, $CD71^{+}$ cells also expressed either CD45 or GPA, but minimal co-expression of CD45 & GPA was observed. As a result of this, almost all of the $CD71^{+}$ cells in the GPA- fractions also expressed CD45, making the target cell population $CD45^{+}CD71^{+}GPA^{-}$ erythroid progenitors, which made up between 79.7-88.7% of the GPA-CD71+ fraction.

While CD71 enrichment was successful, increasing the $CD71^{+}$ population in the GPA-CD71+ fraction to 84.9%, it appears as though a large number of $CD45^{+}CD71^{+}$ cells did not bind the column, and made up an average of 57.9% of cells in the GPA-CD71- fraction, rising to as high as 79.8% for Sample 3. It is unclear what caused this, since given the low cell numbers involved it is unlikely that the column was saturated.

It is possible that CD71 is expressed at low levels on the cells in the GPA-CD71- fraction, which could be due to their being at an earlier stage of development. $CD71^{LOW}$ cells may not have reached the magnetic bead binding threshold for effective enrichment, but might still have been detectable by flow cytometry.

While there is some observable variation in the intensity of the CD71 staining between the two GPA- fractions, it is not clear whether this would be sufficient to account for different binding affinities to the column (Figure 3.15). Interestingly, two distinct CD45+ populations with different CD45 staining intensities are observed in both the GPA-CD71- and GPA-CD71+ fractions. CD71 expression does not appear to be specifically associated with either $CD45^{HIGH}$ or $CD45^{LOW}$ in either fraction, but higher intensity of CD71 expression does appear to be more prevalent on cells with higher levels of CD45. Neither cell population stained positive for GPA.

It has also been suggested that the retention of a significant population of CD71 expressing cells in the CD71- fractions may be due to the fact that a different antibody was used for the flow cytometric analysis than that used for the purification. The two different antibody clones recognise different epitopes, and since CD71 is glycosylated, it is possible that the specificity for CD71 glycosylation variants is not the same for the two antibodies[403]. This could lead to specific isoforms of CD71 identified in the flow cytometric analysis not being bound to the column during the purification step, which would explain the observed results. This could be tested by using alternative antibodies for the flow cytometric analyses, however it is not advisable to use the same antibody for the two different experiments, since if binding sites are saturated during the purification process, then the fluorescent antibody could be blocked from binding, and cells will appear to be CD71 negative.

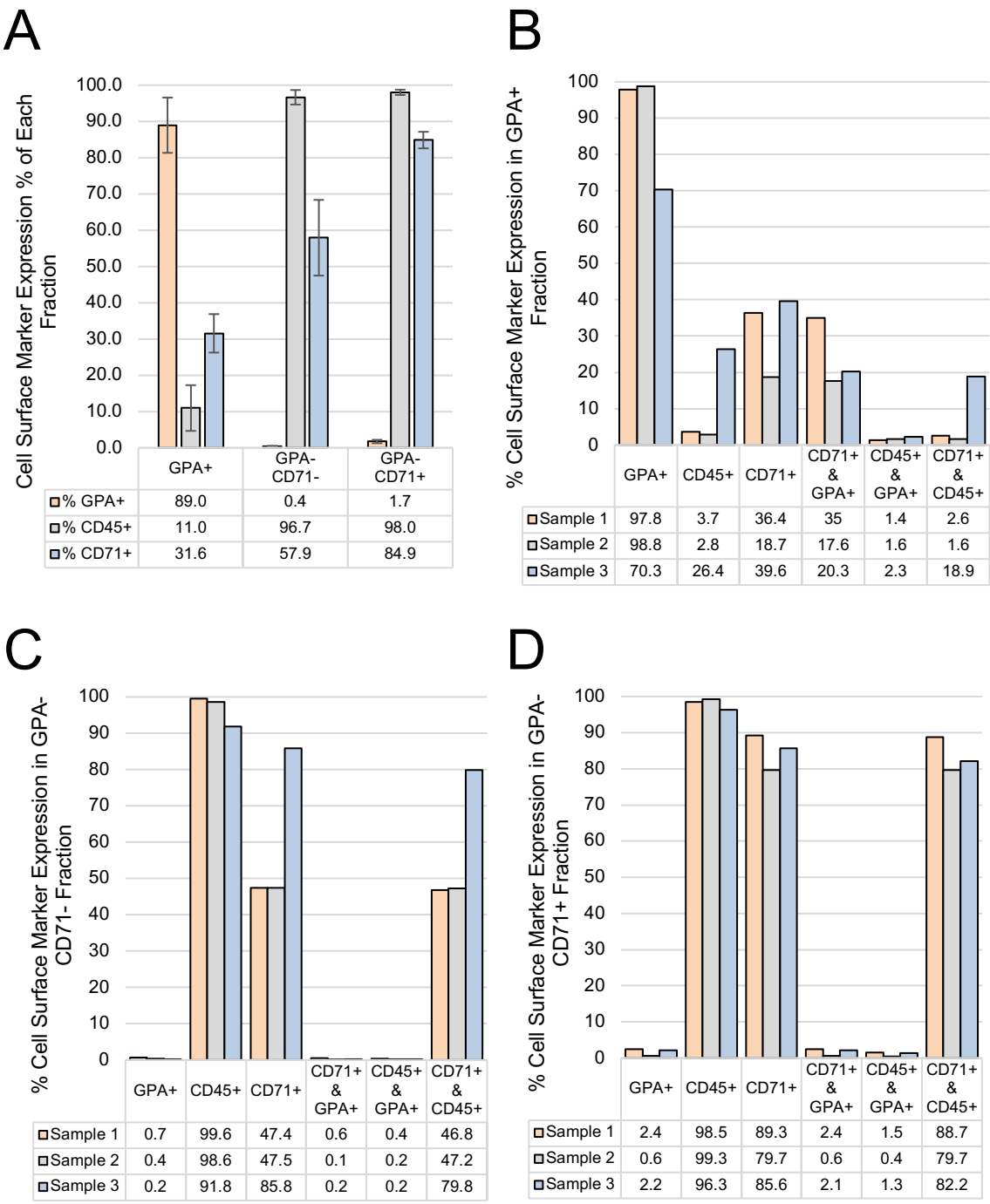
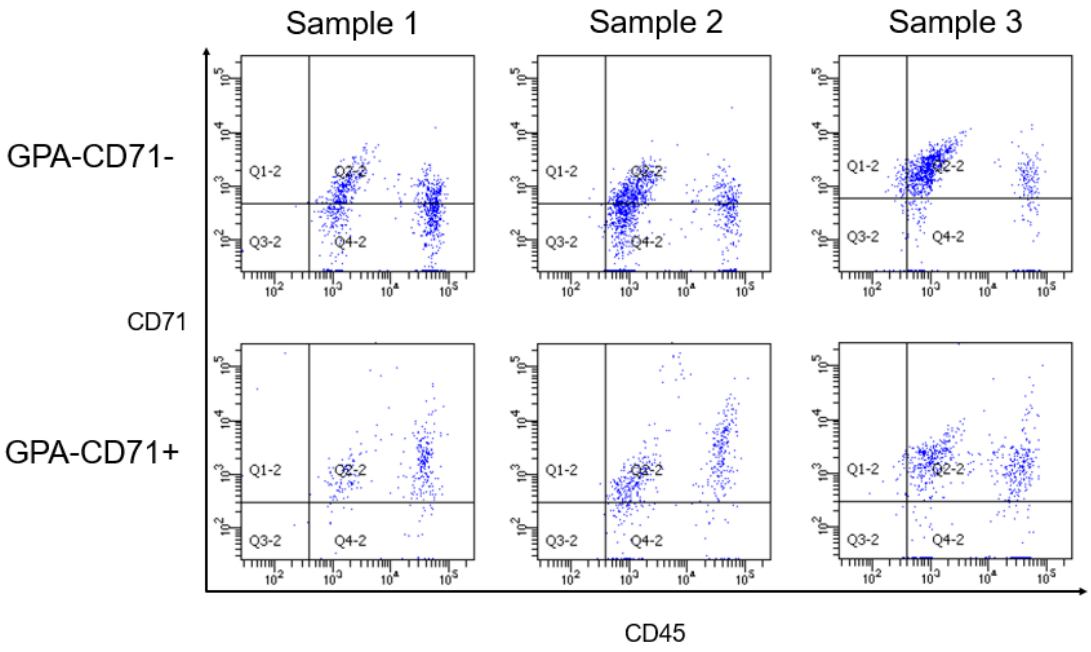


Figure 3.14: Flow cytometry analyses of the three fractions (GPA+, GPA-CD71- & GPA-CD71+) isolated from three HBSS patient samples by GPA depletion and subsequent CD71 enrichment. Sample 2 was receiving HU treatment. A – Mean percentage of cells positive for GPA, CD45 & CD71. Error bars represent standard error. GPA^+ cells were successfully depleted, making up 89.0% of the GPA+ fraction and 0.4% and 1.7% of the GPA- fractions. $CD71^+$ cells were high in both the CD71- and CD71+ fractions, although higher in the enriched fraction, at 84.9%. B, C & D show individual expression as well as co-expression of markers for cells in each of the three fractions: B – GPA+. C – GPA-CD71-. D – GPA-CD71+. $CD45^+$ cells made almost all of the GPA- fractions, and as was observed previously, very little co-expression of GPA and CD45 was observed. $CD71^+$ cells made up 84.9% of the GPA-CD71+ fraction, with 83.5% co-expressing CD45.

Approximately 400ng of DNA was extracted from the GPA-CD71+ fractions of each of the three samples tested (Figure 3.15). While this is still below the 500ng recommended for genome-wide

DNA methylation analysis using the Infinium Infinium® HumanMethylation450 BeadChip, it is a much higher yield than was obtained using any of the other techniques, and would hopefully be sufficient to generate informative data on patterns of DNA methylation in these cells[395].

A



B

	Sample 1	*Sample 2*	*Sample 3*
Cell number	150,000	200,000	60,000
Extraction Method	Q-Micro	Q-Micro	Q-Micro
DNA Concentration (ng/µl)	8.28	8.50	8.56
DNA Total Yield (µg)	0.41	0.43	0.43

Figure 3.15: Analysis of samples after depletion of GPA$^+$ cells and enrichment for CD71$^+$ cells. A – Flow cytometry plots for CD71 and CD45, comparing the GPA-CD71- and GPA-CD71+ fractions for all three samples tested. Intensity of CD71 is higher for some cells in the fraction enriched for CD71. Two distinct CD45$^+$CD71$^+$ populations are visible, distinguishable by high or low CD45 expression. B – Table summarising the DNA extracted from the GPA-CD71+ fractions of the three samples. Q-Micro – Qiagen QiaAMP DNA Micro Kit. Very low cell numbers were isolated, but total DNA yield is in the region of 400ng for all three samples, just below the 500ng recommended for DNA methylation analysis[395].

3.6 Summary of Erythroid Progenitor Isolation Results

The identification of a non-invasive technique that allows reliable and reproducible study of erythroblasts from the peripheral blood of SCA patients will be highly valuable, and will allow longitudinal studies of patients undergoing treatment. This will be especially useful for investigation into the mechanism of action of HU, and how response varies between patients. As was described in 1.5.4, the mechanism by which HU results in the upregulation of HbF in SCA patients is not fully understood, but appears to involve targeted regulatory changes in key erythroid transcription factors, such as MYB, BCL11A, KLF1 and TAL1, rather than merely as a result of increased stress erythropoiesis in response to the cytotoxic effect of the drug[81,239,240,246,247].

Since this change in HbF levels has also been observed in response to treatment with 5-azacytidine, a potent inhibitor of DNA methylation which also has cytotoxic effects, we hypothesised that there could be a role for epigenetic regulation, and specifically DNA methylation in the mechanism of action of HU[254,256].

The importance of DNA methylation at the γ-globin promoter in the silencing of γ-globin is well established, and therefore presents itself as an obvious target for DNA demethylation[70,101,102,104]. This has previously been investigated by Walker *et al*. who found that HbF induction in response to HU was not accompanied by hypomethylation at the γ-globin promoter[245]. However, the absence of changes in DNA methylation at one promoter is not sufficient to rule out the possibility of it playing an important role in another part of the complex regulatory pathway. We hypothesise that performing genome-wide methylation analysis as part of a longitudinal study in SCA patients undergoing HU therapy will provide important insight into the global epigenetic changes that occur, and when coupled with RNA-seq will inform on specific regulatory changes that may be causing the observed increase in HbF expression.

However, the results presented in this chapter demonstrate that using the *in vitro* culturing technique as we initially proposed, we were unable to reliably obtain an erythroid progenitor population for analysis in longitudinal studies. As a result of this, we were unable to investigate the effect that HU treatment has on the epigenome of SCA patients. The combined unreliability of the culture technique along with the difficulty in obtaining longitudinal blood samples from patients on HU meant that this approach would not have provided sufficient data points for a

statistically reliable study. Therefore we took the decision to change tack at this point in the research plan.

We demonstrated that a $CD71^{+}GPA^{+}$ cell population exclusive to the peripheral blood of SCA patients, that we had thought represented late stage nucleated progenitors, actually represented enucleated reticulocytes, presumably as a result of increased stress erythropoiesis in these patients. Through further investigation of erythroid progenitors at an earlier developmental stage, we identified a $CD45^{+}CD71^{+}GPA^{-}$ nucleated cell population that we were able to successfully isolate.

As such, we propose a method of GPA depletion, followed by CD71 enrichment to obtain an early stage erythroid progenitor population from small volumes of peripheral blood from SCA patients. DNA extracted from this population should be sufficient to perform genome-wide DNA methylation analyses, and this technique will be used to conduct future longitudinal studies on the effect of drug treatment on the methylome of SCA patients, should they become accessible in sufficient numbers.

Chapter 4 Results: Whole Exome Sequencing Analysis of Sickle Cell Anaemia Patients

4.1 WES Study Rationale

This work was undertaken to identify novel genetic modifiers of SCA phenotype severity. We hypothesised that a large number of these modifiers remain undetected, and could cause the large amount of variation in disease severity that is currently unaccounted for.

As described previously, there is a huge variation in the severity of phenotype of SCA patients, despite the fact that the disease is considered to be a monogenic disorder, manifesting in a simple Mendelian recessive manner. While in simplistic terms this is the case, there have also been shown to be a variety of contributing factors that influence how the disease is presented. Heterozygous carriers of the HbS allele are not always asymptomatic, and some homozygous patients have a phenotype mild enough to be considered healthy.

A range of genetic modifiers have been identified, and these are discussed in 1.6. However, most of the variation observed remains unaccounted for, and we hypothesised that additional genetic factors modifying the phenotype remain to be discovered.

WES is a powerful and cost-effective tool for identifying genetic variants in case-control studies, and although not able to detect variants in the non-coding regions e.g. mutations that disrupt long range promoter-enhancer interactions, it has been estimated that 85% of the known disease casing mutations fall into the 1% of the genome that is covered by exome capture kits[404,405].

In this study, WES is used to interrogate groups of phenotypically mild and severe SCA patients, with the aim of identifying novel genetic modifiers of the disease. We sequenced 21 mild and 5 severe SCA exomes from a collection at King's College Hospital (KCH), and downloaded 651 publicly available exomes from dbGaP (phs000691.v2.p1). This included 132, 139 and 140 exomes from patients recruited to the SWiTCH, TWiTCH and HUSTLE clinical trials, respectively. Three different strategies were used to analyse these datasets to achieve our aims:

1. Analysis 1: We aimed to identify individual variants that are protective of the severe SCA phenotype. This was achieved by identifying SNPs with high frequency in the mild patient group that were absent from the severe group, and applying a series of filtering

criteria to remove variants that are less likely to affect the pathophysiology of the disease. Each of these filtering steps introduces an inherent bias into the candidate variant list generated, and are discussed in detail in 4.3.1.

2. Analysis 1: Using the candidate variant lists generated, a gene burden test was performed. This aimed to identify genes that contained candidate variants in as many of the mild patients as possible.

3. Analyses 2 & 3: A series of purely statistical analyses were performed. Using Fisher's Exact Tests to identify variant enrichment between either the mild and severe SCA patient groups, or between the SWiTCH and HUSTLE groups. Unlike in the previous tests, the only bias introduced is the selection for variants that occur in the coding region.

The implementation of these aims is summarised in Figure 4.1, which demonstrates how the exome data generated from the KCH mild and severe cohorts, as well as the US SCA datasets SWiTCH, TWiTCH and HUSTLE were used to identify candidate modifier variants for the SCA disease phenotype.

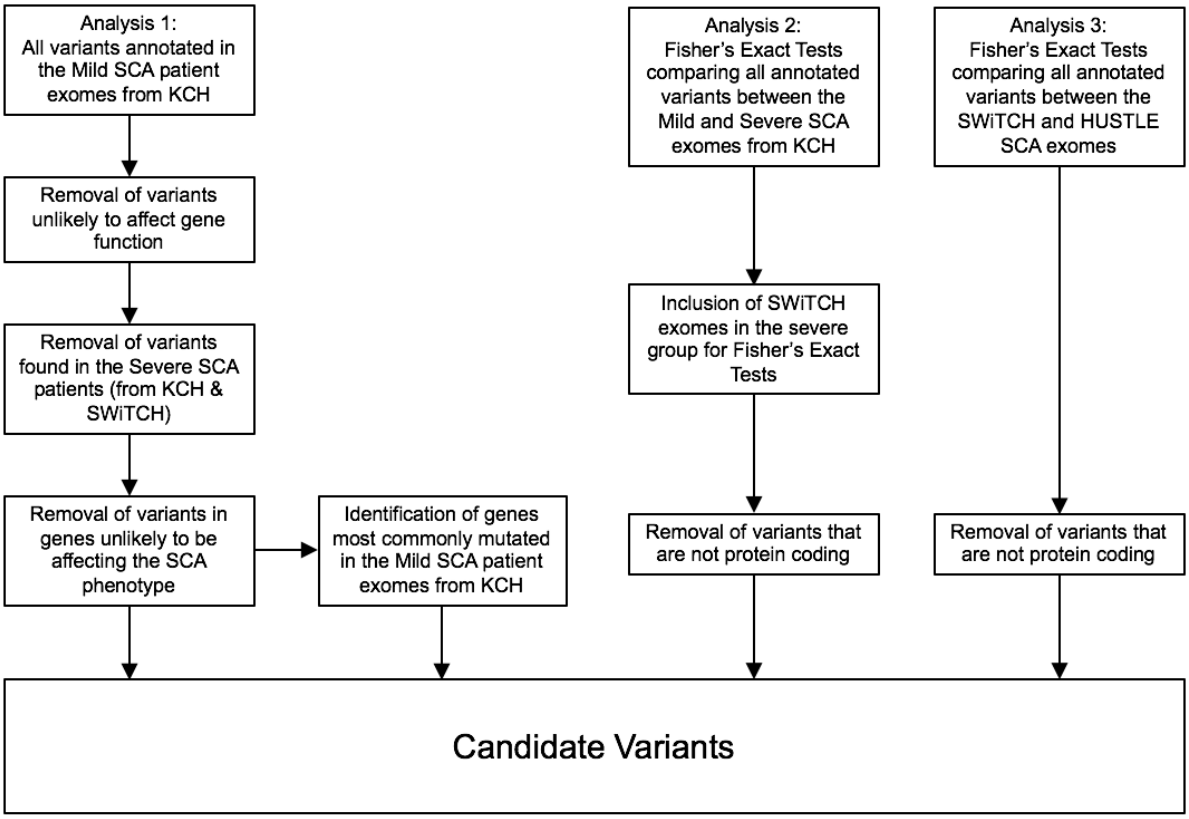


Figure 4.1: Flow diagram outlining the three different analyses performed in this chapter in order to identify candidate genetic modifiers of SCA. Analysis 1 is presented in 4.3 and 4.4, with a detailed description of the various filtering steps provided in 4.3.1. Analysis 2 is presented in 4.5.2, and Analysis 3 in 4.5.3.

4.1.1 Stratification by Clinical Phenotypes

When WES is combined with detailed clinical information, it allows for powerful in depth analysis of genotype:phenotype relationships within patient cohorts. In studies with sufficient numbers of patients, the patient population can be broken down into sub-populations, defined by highly specific clinical characteristics. In the case of severe SCA patients, this could include frequency of sickle cell pain crises, sickle-related organ damage, and different types of stroke.

In the case of stroke, the inclusion criteria used in this study simply required overt clinical stroke, as opposed to silent infarctions, which do not present with strong clinical symptoms, and are diagnosed retrospectively, such as by MRI scans[406]. With a large enough sample size, overt clinical stroke could be stratified into sub-groups of ischaemic or haemorrhagic stroke, which have different patterns of occurrence between different SCA patient age groups[407,408]. Ischaemic strokes occur as a result of vaso-occlusion or narrowing of blood vessels in the brain, and a study has shown that SCA patients are at the highest risk below the age of 19, and above the age of 30[407]. Conversely, haemorrhagic strokes are caused by rupturing of blood vessels in

the brain, and SCA patients were shown to be most at risk between the ages of 20-30[407,409]. Due to the difference between these types of strokes, and the difference in the age at which they occur, it is likely that different pathophysiological pathways could be involved, with different underlying genetic modifiers. Stratifying these two groups would therefore allow more sensitive analyses. Of the four SCA patients recruited from King's College Hospital that had experienced a stroke before the age of 18 (Table 4.1), each of these included at least one ischaemic stroke, and two had subsequently had haemorrhagic strokes.

In this project, due to the limited size of both the mild and severe SCA groups, and the fact that no individual phenotype information was available for the US SCA exome dataset, stratification of the populations on specific symptoms was not performed. In the future, if the sample size of the study is substantially expanded, this could provide valuable insight into the underlying mechanisms that result in the wide range of symptoms presented by SCA patients. While it would be more complicated to incorporate into the variant filtering pipeline performed for Analysis 1, it would be relatively straightforward to perform this sort of stratification for Analyses 2 & 3, which would allow identification of variants not just associated with mild or severe forms of the disease, but also of variants that associate with specific symptoms.

4.2 SCA Patient Data Summary

4.2.1 SCA Patients from King's College Hospital

DNA samples from 26 SCA patients stored at King's College Hospital were Whole Exome Sequenced for this study. Patients with the HbSS genotype were selected to avoid any phenotypic variation associated with other forms of SCD. Patients were selected from the extreme ends of the phenotypic range, with the aim of comparing severe and mild patients to identify novel genetic variants influencing the severity of the SCA phenotype. Of the 26 patients, five were categorised as severe, and 21 as mild. Fewer severe patients were selected than mild, since there are data from a US WES study that is publicly available through dbGaP (phs000691.v2.p1), which included severe SCA patients[376].

4.2.1.1 Severe Patients

Severe patients were selected based on the severity of clinical symptoms presented at a young age, these are summarised in Table 4.1. One major indicator used was having an ischaemic stroke in childhood, which was experienced by four out of the five patients selected, and in one extreme case three strokes had occurred by age 6. One patient had not experienced stroke, but presented a wide range of other symptoms, and while these symptoms are not rare complications of SCA (many can be observed in the mild patients in Table 4.2), they are often associated with older patients, and it is rare to experience all of them at a young age. This patient is also undergoing HU therapy, which may have prevented stroke from occurring.

Patient ID	*Age*	*Sex*	*Treatment*	*Genotype*	*Alpha*	*Mean HbF% (Age)*	*Ischaemic Stroke (Age)*	*Other Symptoms*
GMKH 001	30	F	Blood Transfusion	HbSS	αα/α-	Not Known	13, 23*	Retinopathy, Pulmonary Hypertension, Gallstones, Fe Overload
GMKH 042	29	M	Blood Transfusion	HbSS	αα/αα	Not Known	3, 4, 6, 20	Abnormal Proteinuria, Acute Chest Syndrome, Parvovirus Infection, Fe Overload
GMKH 063	25	F	Blood Transfusion	HbSS	αα/αα	6.9 (11)	18	Gall Bladder Removed, Acute Chest Syndrome, Osteoarthritis, Hypertension
GMKH 234	31	F	Blood Transfusion	HbSS	αα/α-	Not Known	8, 23*	Abnormal Proteinuria, Gall Bladder Removed, Fe Overload
GMKH 249	24	M	Hydroxyurea	HbSS	αα/αα	3.1 (22)	None	Abnormal Proteinuria, Pulmonary Hypertension, Acute Chest Syndrome, Leg Ulcers, Hyperhaemolysis, Deep Vein Thrombosis, Osteomyelitis

Table 4.1: Patient data for the 5 severe phenotype SCA patients that were sequenced. Samples GMKH 001, 042, 063 & 234 all had a stroke at ≤18 years old. Patient GMKH 249 did not have a history of stroke, but was classified as severe due to the severity of other symptoms experienced at a young age. Stroke refers to ischaemic stroke, * indicates haemorrhagic.

Two of the severe patients have the silent carrier genotype of α-thalassaemia, this is interesting since it has been shown that the carrier state can contribute to a reduction of severity of some SCA complications[294–296,298]. HbF% is not reported for all patients on blood transfusions, since contamination with the donor blood prevents accurate detection while undergoing regular treatment.

4.2.1.2 Mild Patients

Mild patients were selected on the basis of not having experienced a stroke or any other severe complication of SCA by the age of 30, the clinical information regarding these patients is summarised in Table 4.2.

It was discovered that two of these patients had previously been found to have concurrent α-thalassaemia. Although these were only αα/α- & α-/α- genotypes, representing silent carrier and α-thalassaemia trait respectively, they would still be expected to modify the SCA phenotype. Due to having an already identified causative variant, these patients were excluded from the study.

One patient was heterozygous for the sickle cell mutation and a β^0-thalassaemia allele, and was included due to this genotype's similarity to HbSS in terms of both disease pathology and clinical severity. Despite being heterozygous, HbS is the only form of haemoglobin expressed in HbS/β^0 patients (as described in 1.2.3.3).

While mild patients with concurrent α-thalassaemia were excluded, patients with abnormally high percentage HbF were not. This is because the diagnosis of α-thalassaemia is confirmed by genetic testing to identify the genotype, and therefore has a known cause. In the case of HPFH disorders however, HbF is measured by High Performance Liquid Chromatography (HPLC) as a percentage of total haemoglobin, and so is measured as a phenotype, the cause of which is not necessarily explained. As described in 1.6.1 there are a variety of genetic factors known to influence HbF levels, and the SCA patient collection at King's College Hospital has previously been screened for these common variants. Only the patients with no known cause for HPFH were included in this study.

Patient ID	*Age*	*Sex*	*Treatment*	*Genotype*	*Alpha*	*Mean HbF% (Age)*	*Stroke (Age)*	*Other Symptoms*
GMKH 169	33	F	None	HbSS	αα/αα	20.0 (20)	None	Abnormal Proteinuria
GMKH 171	38	F	None	HbSS	αα/αα	8.4 (25)	None	Gall Bladder Removed, Spleen Removed
GMKH 095	52	F	Intermittent Blood Transfusions	HbSS	αα/αα	6.4 (41)	None	Retinopathy, Abnormal Proteinuria, Gall Bladder Removed, Asymptomatic Acute Chest Syndrome
SCD 178	45	F	None	HbSS	αα/αα	17.7 (32)	None	Asymptomatic Acute Chest Syndrome
GMKH 138	55	F	None	HbSS	αα/αα	19.3 (51)	None	Abnormal Proteinuria, Gall Bladder Removed, Avascular Necrosis, Carcinoid, Deep Vein Thrombosis
SCD 213	74	F	None	HbSS	αα/αα	29.5 (62)	None	Abnormal Proteinuria, Osteoarthritis
SCD 215	55	F	None	HbS/ β°	αα/αα	9.2 (49)	None	Retinopathy, Abnormal Proteinuria
GMKH 016	43	F	None	HbSS	αα/αα	8.3 (28)	None	High Proteinuria, Asymptomatic Acute Chest Syndrome, Hepatitis B
HFKH 063	45	F	None	HbSS	αα/αα	10.5 (33)	None	High Proteinuria, Small Distal ICA Aneurisms
GMKH 056	62	F	None	HbSS	αα/αα	1.3 (50)	None	NK, Gallbladder Removed, Avascular Necrosis, Asymptomatic Acute Chest Syndrome, Thromboembolism
SCD 278	54	M	None	HbSS	αα/αα	13.4 (45)	None	Retinopathy, Abnormal Proteinuria, Avascular Necrosis, Priapism
SCD 146	54	F	None	HbSS	αα/αα	16.3 (53)	None	Gall Bladder Removed, Avascular Necrosis, Asymptomatic Acute Chest Syndrome
GMKH 317	59	M	None	HbSS	αα/αα	1.8 (57)	None	Retinopathy, Abnormal Proteinuria, Gall Bladder Removed, Mild Acute Chest Syndrome, Leg Ulcers
GMKH 052	49	F	Hydroxyurea	HbSS	αα/αα	15.2 (42)	None	NK, Gallstones
GMKH 084	38	F	None	HbSS	αα/α-	9.5 (32)	None	Pes Planus
GMKH 143	62	F	Hydroxyurea	HbSS	αα/αα	22.3 (56)	None	Abnormal Proteinuria, Mild Pulmonary Hypertension, Acute Chest Syndrome, Pulmonary Embolism, Abdominal Hernia, Hypertension, Pes Planus
GMKH 175	69	F	None	HbSS	α-/α-	7.9 (64)	None	Abnormal Proteinuria, Borderline Pulmonary Hypertension, Gall Bladder Removed, Hypertension, Cataracts, Hepatitis B
GMKH 290	64	F	None	HbSS	αα/αα	11.4 (59)	None	Borderline Pulmonary Hypertension, Gall Bladder Removed
GMKH 036	63	M	None	HbSS	αα/αα	7.1 (58)	None	Abnormal Proteinuria, Borderline Pulmonary Hypertension, Gall Bladder Removed, Avascular Necrosis
GMKH 179	69	M	None	HbSS	αα/αα	20.3 (63)	None	Prostate Cancer, Osteomyelitis, Inguinal Hernia
SCD 131	52	F	None	HbSS	αα/αα	16.1 (45)	None	Bilateral Carpel Tunnel Syndrome

Table 4.2: Patient information for the 21 mild phenotype SCA patients that were sequenced. None of the samples had had a stroke by age 33, and the majority are not on any form of treatment. Patients GMKH 084 & GMKH 175 both have concurrent α-thalassaemia and were excluded. Patient SCD 215 was heterozygous for Sickle Cell Trait and β^0-thalassaemia, which is phenotypically similar to HbSS.

4.2.2 SCA Exome Data from dbGaP

651 SCA patient WES datasets were acquired from dbGaP (phs000691.v2.p1). Data for two of these patients were unable to be correctly downloaded, despite multiple attempts and so are not included. A summary of the source of the remaining 649 samples is shown in Figure 4.2.

411 of these patients were recruited from one of three clinical studies investigating the outcome of HU treatment in SCA patients in the USA. Other than sex, age at start of HU treatment, and HbF percentage before and after treatment, no individual clinical information is available for these patients.

For the patients that can be identified as being recruited through one of the clinical trials (HUSTLE, SWiTCH[251] & TWiTCH[252]), it is possible to categorise them based on the specific inclusion criteria for each trial. These inclusion criteria are summarised in Table 4.3.

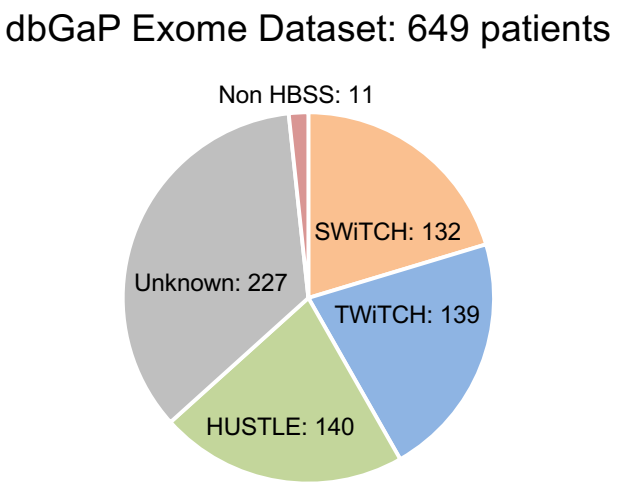


Figure 4.2: Summary of 649 SCA exomes downloaded from dbGaP (phs000691.v2.p1). Samples were checked for the SCA mutation (rs334), 10 were found to be heterozygous, and 1 found to be homozygous for the wild type, these samples were excluded from further analyses. The majority of patients (411) were recruited from one of the three clinical trials – HUSTLE, SWiTCH or TWiTCH.

	HUSTLE	*SWiTCH*	*TWiTCH*
Age (years)	<30	5 - 18	4 – 15
Sex	M + F	M + F	M + F
Clinical Criteria	Patients currently receiving or about to receive Hydroxyurea therapy*.	Overt clinical stroke >1 year old, documented by CT or MRI. >18 months of chronic erythrocyte transfusions since primary stroke. Transfusional iron overload. Adequate monthly erythrocyte transfusions with average HbS <45% in the 6 months prior to study entry.	Documented index (pre-treatment) abnormally high TCD Velocity by Transcranial Doppler ultrasonography. >12 months of chronic erythrocyte transfusions since abnormal TCD. Adequate monthly erythrocyte transfusions with average HbS <45% in the 6 months prior to study entry.
Exclusion Criteria	n/a	Inability to receive RBC transfusion therapy. Inability to take daily oral hydroxyurea. Clinical and laboratory evidence of hypersplenism. A sibling enrolled in SWiTCH.	Overt clinical stroke or TIA. Known severe vasculopathy or moyamoya disease on brain MRA. Inability to receive chronic RBC transfusion therapy. Inability to take daily oral hydroxyurea. Clinical and laboratory evidence of hypersplenism. A sibling enrolled in TWiTCH.
Non-SCA Controls?	No	No	No
Patients Recruited	260	134	159

Table 4.3: Table summarising recruitment criteria for the three clinical studies – HUSTLE, SWiTCH & TWiTCH. Information is obtained from clinicaltrials.gov website, and is correct as of November 2016. * - Inclusion criteria for HUSTLE only require patients to be taking HU, additional information on the criteria for prescribing treatment at St. Jude's Children's Hospital, the trial centre, is described by Nottage *et al.* 2014[205].

4.2.2.1 Stroke with Transfusions Changing to Hydroxyurea (SWiTCH)

SWiTCH was a study of SCA patients that had suffered a stroke at a young age and were receiving regular blood transfusions. The trial investigated the benefits of switching these patients to HU treatment in an attempt to avoid the side effects of long-term blood transfusion therapy. The trial was ultimately stopped when liver iron levels, one of the primary outcomes, was found not to be improved in the HU treatment arm, and an increase in the frequency of adverse effects was observed[222,250,251].

Based on the study inclusion criteria shown in Table 4.3, it is clear that all patients recruited through SWiTCH must have had at least one stroke before the age of 17.5 (to allow at least 18 months of blood transfusions before the age of 19), this is in line with the criteria used for definition of severe patients that were sequenced from King's College Hospital. As such the 132 exomes from the SWiTCH study are included as part of the severe dataset for our analyses.

4.2.2.2 Transcranial Doppler (TCD) With Transfusions Changing to Hydroxyurea (TWiTCH)

TWiTCH was a study of SCA patients that had not suffered a stroke, but were identified as having abnormally high Transcranial Doppler velocities (TCDv), one of the key indicators for risk of stroke, and had been receiving regular transfusions as a preventative treatment[215,216]. The trial investigated the benefits of switching these patients to HU therapy, to avoid the side effects of long-term blood transfusion therapy, similar to the SWiTCH trial. TWiTCH demonstrated that the HU treatment arm was non-inferior to blood transfusions in terms of reducing TCDv, and showed a significant decrease in liver iron levels compared to those in the transfusion treatment arm, which increased[252].

Although abnormal TCDv is an indicator for risk of stroke, and regular blood transfusions would be expected to prevent stroke occurring, it is not possible to categorise all of these patients as severe by the same criteria used for the patients recruited from King's College Hospital. Additionally, of the 159 patients recruited, only 121 passed the initial screening stages, meaning that up to 38 of the 139 downloaded TWiTCH exomes didn't fully meet the inclusion criteria[252]. Although, 29 out of the 38 were either excluded for severe vasculopathy, or chose to withdraw, and would still have the same clinical definition as the included patients[252].

Due to concerns about the clinical definition of this 'high risk' group, the 139 exomes from the TWiTCH study are only used in combination with the severe dataset in some parallel analyses, with a lower threshold for definition of severe patients.

4.2.2.3 Long Term Effects of Hydroxyurea Therapy in Children with Sickle Cell Disease (HUSTLE)

HUSTLE is an on-going observational study, aiming to investigate the long-term effects of HU therapy in SCD. It is not possible to categorise the severity of these patients based on the inclusion criteria alone (Table 4.3), since any patient under the age of 30 being treated with HU could be included. More information is available regarding the clinical criteria for starting patients on HU at St. Jude's Children's Research Hospital, where patients were recruited[205]. These criteria are as follows: 'frequent vaso-occlusive pain, acute chest syndrome, chronic hypoxemia, haemoglobin level less than 7.0 g/dL, HbF less than 8% after 24 months of age, WBC count greater than 20 x 10^9/L, and LDH more than twice the upper limit of normal'[205,410].

Even with this additional information, the clinical severity of the HUSTLE participants is difficult to define without individual information for each patient. As such, the HUSTLE participant data is not included in the comparison of mild and severe SCA patients. However, these data are used as a control group when investigating variant enrichment in the severe SWiTCH group. This analysis is based on the assumption that HUSTLE is representative of the SCA population as a whole, and is not strongly biased towards severity of phenotype.

4.3 Analysis 1: Identification of Coding SNPs Protective of the Severe SCA Phenotype

In order to identify genetic variants that are protective of the severe phenotype of SCA, coding variants that are present in the mild group, but completely absent from the severe group were investigated. In this model, any variants identified are assumed to be completely protective, and both recessive and dominant models are taken into account.

4.3.1 Variant Filtering

This section describes the various filtering criteria used to generate a list of candidate variants and the assumptions that accompany these criteria. An overview of the filtering process is demonstrated in Figure 4.4. Figure 4.3 shows the total number of variants in the mild group, and the proportions of these based on type. 48% of all variants are intergenic, and 45% intronic, with only 4% representing coding mutations.

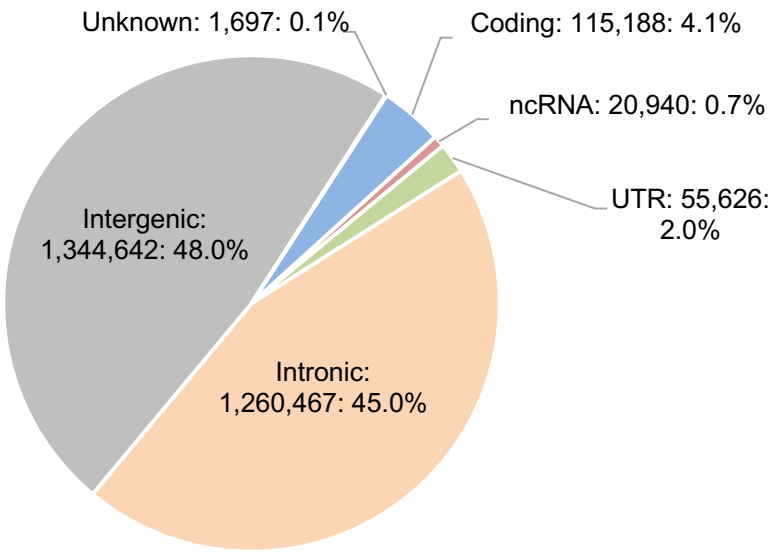


Figure 4.3: Summary of all 2,798,560 variants present in the mild group of patients, grouped by type of variant. Intergenic variants include those annotated as upstream or downstream. Coding variants also include those annotated as splicing variants. UTR – Untranslated Region. 93% of all annotated variants are either intergenic or intronic.

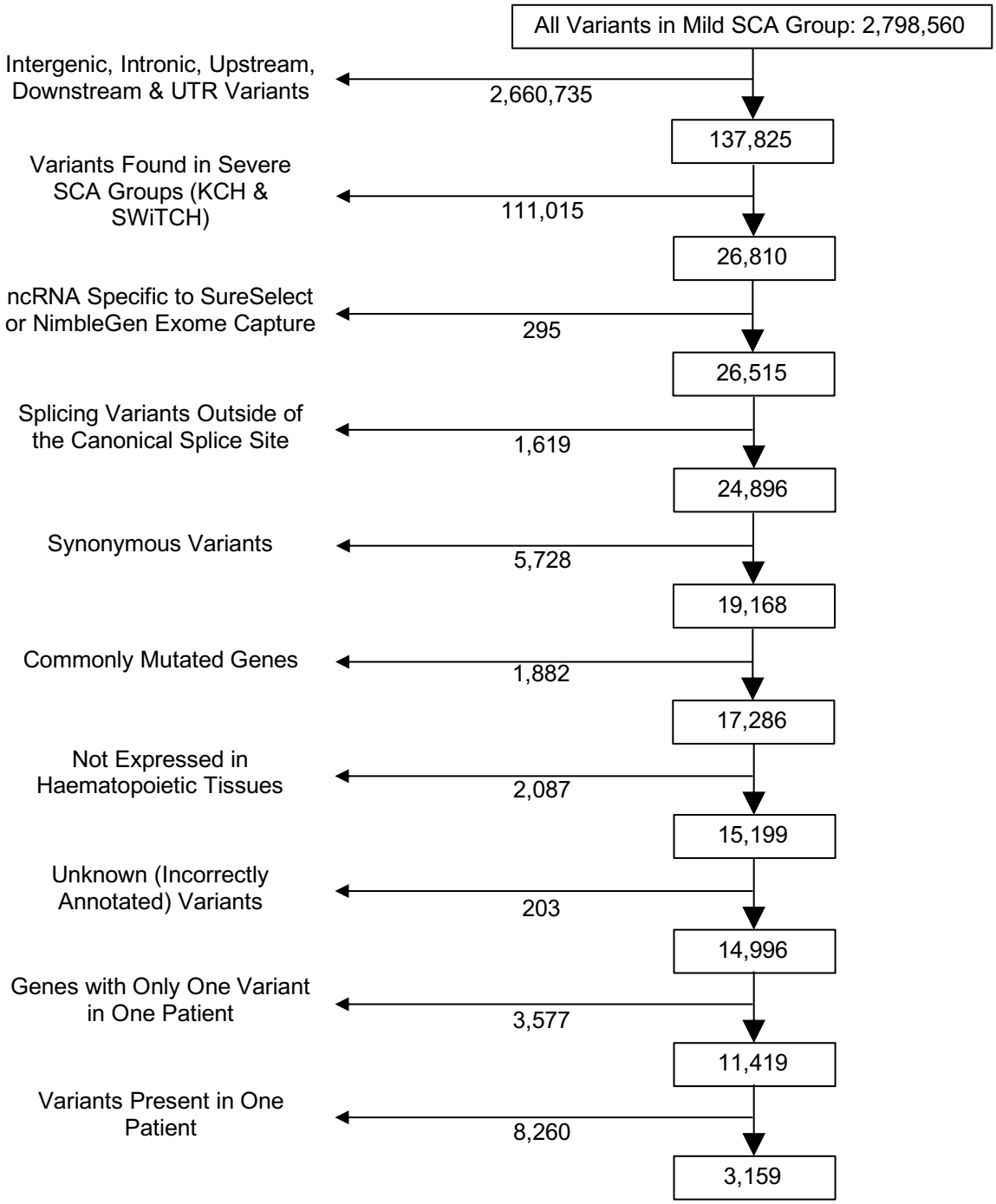


Figure 4.4: Candidate variant filtering pipeline, describing the process of filtering the 2,798,560 variants observed in the mild SCA patient group down to 11,419 for the gene burden analysis, and 3,159 for the individual variant analysis. The full list of 11,419 variants is provided in Appendix 12.

4.3.1.1 Intergenic Variants

When annotating variants with ANNOVAR, intergenic refers to variants found more than 1kb upstream or downstream of gene. While these variants can influence transcriptional regulation through long-range enhancer-promoter interaction, they are situated at the edge of the targeted capture area, and coverage of these regions is inconsistent. Intergenic variants are therefore excluded from the analysis. As well as the intergenic variants, this includes variants <1kb

upstream or downstream of genes, which are annotated by ANNOVAR as 'upstream' or 'downstream'.

4.3.1.2 Non-Coding Variants

The design of this analysis is targeted towards detecting coding variants, since these are the most likely to have a phenotypic effect. Non-coding variants also play an important role in regulation of gene expression, and there are many documented examples of mutations in non-coding regions influencing disease phenotype, including one of the SNPs investigated in Chapter 5. However, since non-coding regions are generally much more tolerant of genetic disruption, it is more difficult to identify variants that are eliciting a true effect. This, combined with the relatively small sample sizes used in this study, means that the analysis will exclude the non-coding variants. As well as the intergenic variants, this includes variants in the untranslated regions (UTRs) of transcripts, and intronic variants.

There are some exceptions to this, and intronic variants near the intron/exon boundary are annotated as splicing variants and retained, similarly UTR variants that span translational initiation or termination codons are also retained. Also variants in ncRNA exons are not excluded.

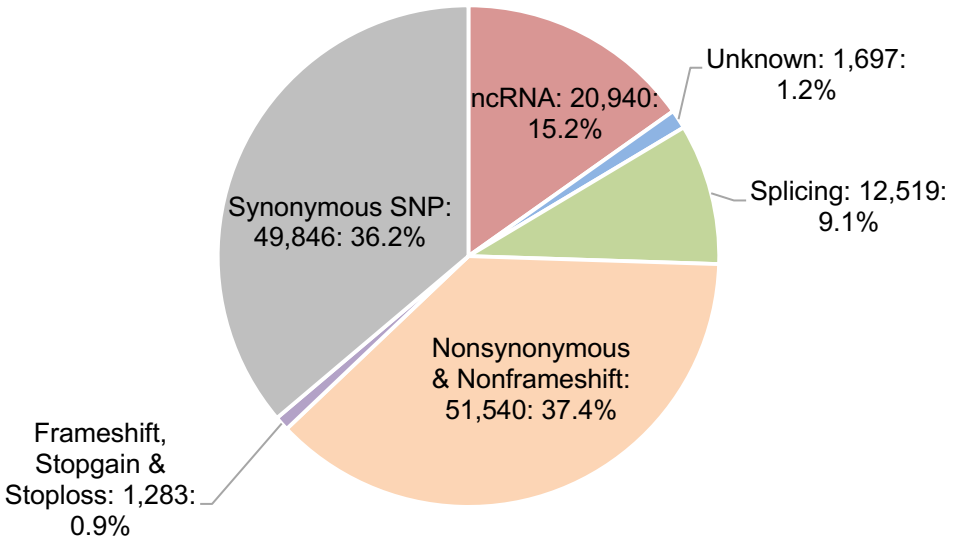


Figure 4.5: Summary of the 137,825 variants present in the mild group after filtering of intergenic and non-coding variants (other than splicing and ncRNA).

Figure 4.5 shows a summary of the candidate variants in the mild group after filtering of non-coding variants. The majority of these are made of both synonymous substitutions (36.2%) and

nonsynonymous substitutions and nonframeshift insertions or deletions (37.4%). These are variants that either have no impact on the amino acid sequence (in the case of synonymous SNPs), or only affect the residue in which the mutation is located, without disrupting the overall sequence or structure of the gene, as is the case in splicing variants, frameshift insertions or deletions and stopgain or stoploss mutations.

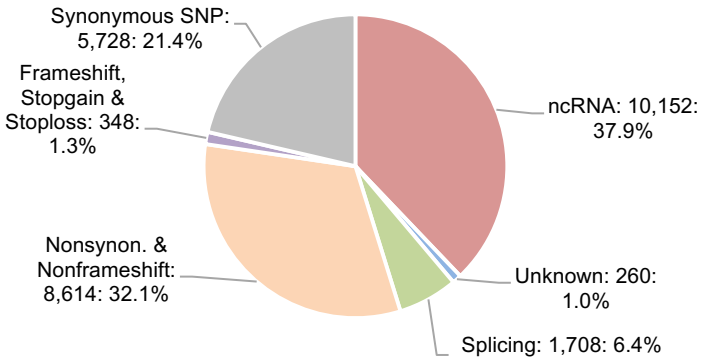
4.3.1.3 Removal of Severe Variants

Variants that are present in one of the severe groups were filtered out, with recessive and dominant models taken into account. For the dominant model, any variant that was only heterozygous in the mild group must be completely absent from the severe group. For the recessive model, if a variant was homozygous in the mild group, it was filtered out if homozygous in the severe group as well, however heterozygous variants in the severe group were tolerated.

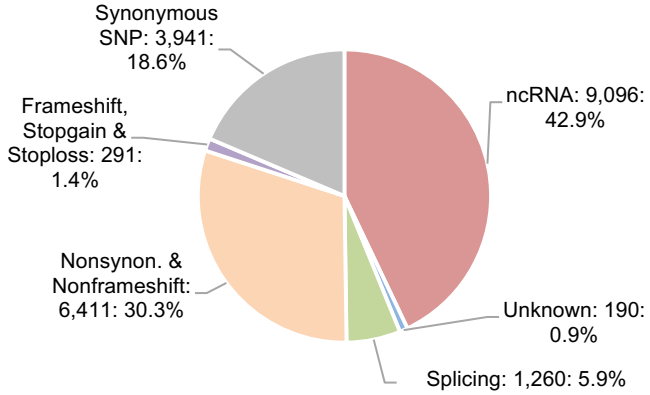
Two separate filtering criteria were used in parallel, one with variants from the severe patients from King's College Hospital and from the SWiTCH trial, and the second additionally excluding variants from the TWiTCH trial, with a less strict definition of severe. The variants after filtering for both of these are summarised in Figure 4.6.

Filtering of variants found in the severe patient group greatly reduced the number of candidate variants, from 137,825 before filtering to 26,810 and 21,189 for the severe and SWiTCH and the severe, SWiTCH and TWiTCH filtered groups respectively. Interestingly the proportion of synonymous SNP variants is reduced from 36.2% of all variants to 21.4% and 18.6%, suggesting that more of these variants are shared between the groups than of the other coding variant types. Presumably this is due to the increased tolerance for variants that don't affect the protein sequence. At the same time the proportion of ncRNA variants increased from 15.2% to 37.9% and 42.9%, suggesting that fewer of these variants are shared, possibly due to the different exome capture kits used, Figure 4.6C summarises the changes in proportion of the different variant types.

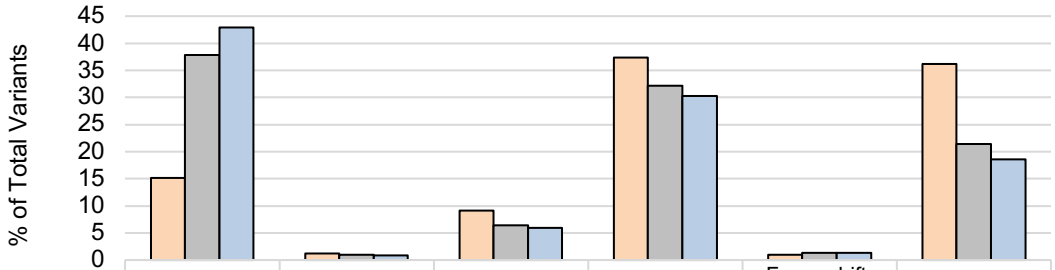
A



B



C



	ncRNA	Unknown	Splicing	Nonsynon. & Nonframeshift	Frameshift, Stopgain & Stoploss	Synonymous SNP
Pre-Filtering	15.2	1.2	9.1	37.4	0.9	36.2
KS	37.9	1.0	6.4	32.1	1.3	21.4
KST	42.9	0.9	5.9	30.3	1.4	18.6

Figure 4.6: Summary of the candidate variants in the mild group after filtering for variants observed in the severe groups. A – Summary of the 26,810 variants after filtering by severe patients from KCH and SWiTCH clinical trial (KS). B – Summary of the 21,189 variants after filtering by severe patients from KCH, SWiTCH and TWiTCH clinical trials (KST). C – Change in proportion of variants for each variant type in A and B compared to before filtering for variants in the severe group (shown in Figure 4.5).

4.3.1.4 Restriction of ncRNA to those targeted by both SureSelect and NimbleGen

Due to the difference in exome capture target areas between the Agilent SureSelect and Roche NimbleGen kits, a list of ncRNA was generated that is present in both the SureSelect and the

NimbleGen exome data. This list was generated using the assumption that each ncRNA will contain at least one variant in one individual for each of the capture kits used. This list is the same as the one used for the analysis in 4.5.2.3 and the complete list is available in Appendix 7. Figure 4.7 summarises the ncRNA filtering step, 4988 ncRNA had variants observed in both the SureSelect and NimbleGen groups, with 336 and 790 respectively expressed exclusively in either group.

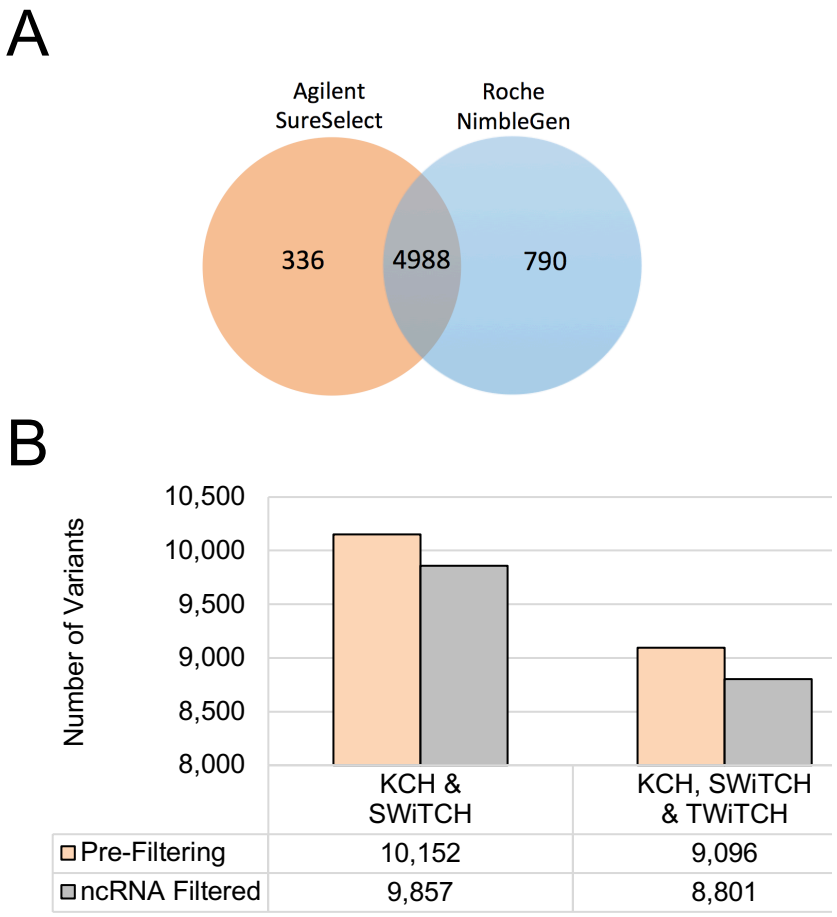


Figure 4.7: Summary of the trimming of the ncRNA dataset to include only variants in ncRNA covered by both the Agilent SureSelect and Roche NimbleGen exome capture kits. A – Summary of the number of ncRNA with annotated variants in each of the exome capture groups. Variants in the 336 ncRNA only present in the SureSelect group were excluded, and only variants in the 4988 that are shared were included in downstream analyses. B – Number of ncRNA variants before and after filtering for each of the candidate variant groups.

4.3.1.5 Splicing Variants

Intronic variants within 10bp of the intron/exon boundary are annotated by ANNOVAR as splicing variants. A degree of variation is permitted within splice site sequences without severely impacting on splicing, there are multiple scoring tools available, but there is still a degree of uncertainty regarding how accurate these tools are[411].

Dinucleotide pairs both upstream and downstream of each exon are strongly conserved in mammalian splice sites, with over 98% featuring a GT downstream and AG upstream[412]. Disruptions of these dinucleotide sites have been shown to have serious effects on splicing efficiency, and when disrupted can lead to exon skipping, alternative exon usage, alternative polyadenylation site usage or intron retention, any of which can significantly impair gene function[413–417]. Due to the strength of the conservation of these canonical dinucleotide splice sites, and the uncertainty of splice site scoring for variants further away, splicing variants were filtered out if situated >2bp up or downstream of the exon. Figure 4.8 shows the number of splicing variants before and after filtering. Filtering based on the variant position within the splice site greatly reduces the number of variants, and confirms that fewer variants are tolerated in the 2bp adjacent to an exon than in the rest of the 10bp sequence. If variants were randomly distributed throughout the 10bp, an 80% reduction would be expected, whereas in Figure 4.8 94.8% and 94.4% reductions are observed respectively for the KCH and SWiTCH and the KCH, SWiTCH and TWiTCH candidate variants.

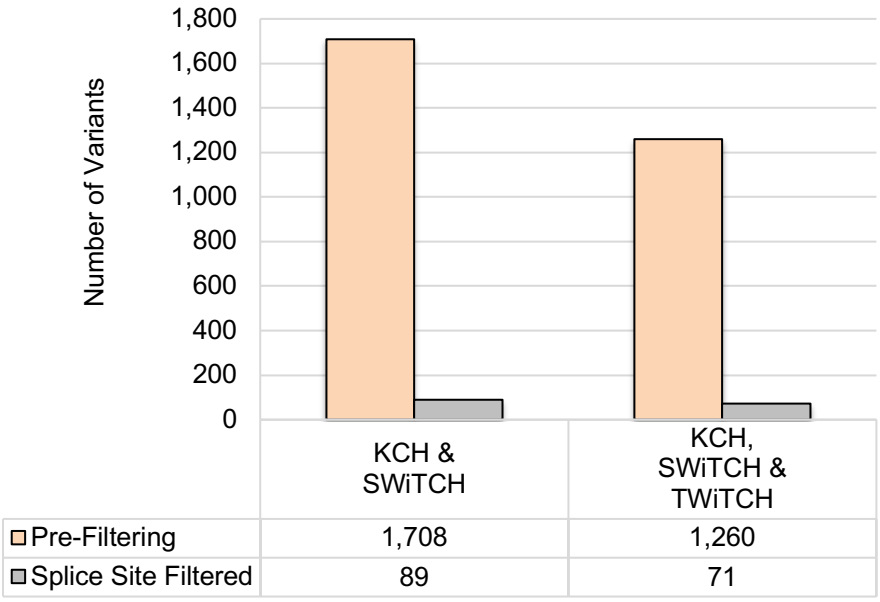


Figure 4.8: Filtering of splicing variants outside of the canonical 2bp splice site, for both the KCH and SWiTCH, and KCH, SWiTCH and TWiTCH filtered candidate variants. Approximately 95% of splicing variants were removed by selecting for 20% of the splice site sequence.

4.3.1.6 Synonymous Variants

Due to redundancies in codon usage, it is possible for nucleotide substitutions to have no effect on the subsequent amino acid sequence. These substitutions are referred to as synonymous, and are effectively silent. Since these variants should have no effect on gene function, they

were filtered out of the candidate variant lists. All 5,728 and 3,941 synonymous SNPs were removed from the KCH and SWiTCH, and the KCH, SWiTCH and TWiTCH filtered candidate variant lists respectively.

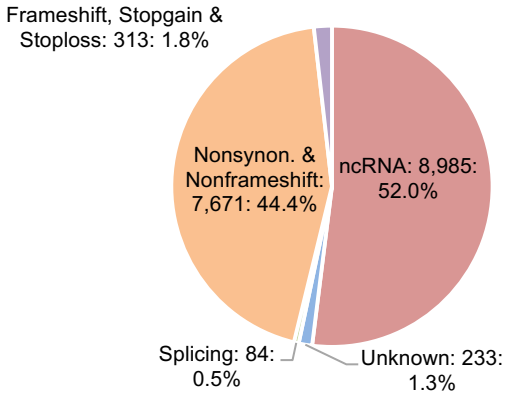
4.3.1.7 Commonly Mutated Genes

Some genes are more commonly mutated than others, and frequently harbour many variants without affecting gene function. This can be due to gene size, if the rate of mutations across coding regions is uniform and considered in the format of SNPs/kb, then it would be expected that for larger genes, more SNPs would be indentified[418]. Additionally, some genes have a higher tolerance for genetic aberration, these typically tend to be genes with a degree of redundancy, or that perform functions that are not particularly dependant on the specific amino acid sequence as much as the overall structure, these include structural proteins and large secreted proteins such as those encoded by the MUC family of mucin genes[419]. These genes are commonly identified in next generation sequencing studies as false positive results.

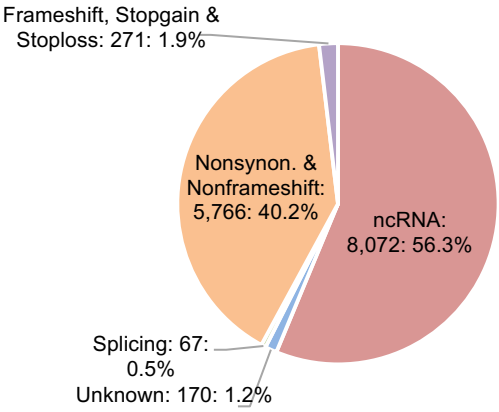
To remove variants found in commonly mutated genes, a list was compiled of genes identified by two published studies[420,421]. The first list is of 2,157 genes and loci from Fuentes *et al*. (2012) that includes 1,580 pseudogenes, 435 genes commonly identified in exome sequencing studies, as well as 142 genes with more than 10 rare nonsynonymous coding variants in more than 3 families from a cohort of 29[420]. The second list is the top 100 genes identified by Shyr *et al.* (2014), in a study investigating **F**requent**L**y mut**A**ted **G**ene**S** (FLAGS), being the genes most commonly found to contain rare coding variants (minor allele frequency <1%) in published exome sequencing studies[421]. The complete list of 2,256 commonly mutated genes is provided in Appendix 10.

A summary of the candidate variants after removal of commonly mutated genes is shown in Figure 4.9. 1,882 and 1,418 variants were removed from the KCH and SWiTCH filtered, and the KCH SWiTCH and TWiTCH filtered groups respectively. The proportion of the variants that occur in ncRNA is increasing, and at this stage make up more than half of the remaining candidate variants. At this filtering step, ncRNA variants made up only 46.3% and 51.4% of those removed.

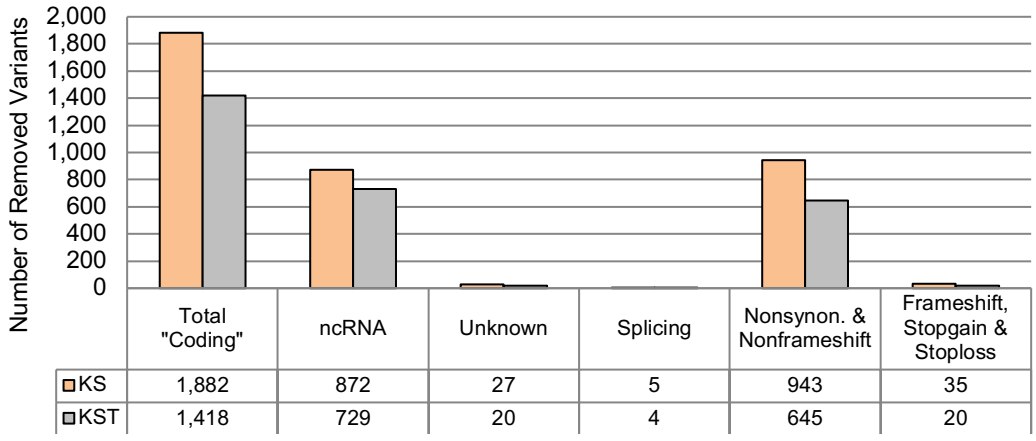
A



B



C



	Total "Coding"	ncRNA	Unknown	Splicing	Nonsynon. & Nonframeshift	Frameshift, Stopgain & Stoploss
KS	1,882	872	27	5	943	35
KST	1,418	729	20	4	645	20

Figure 4.9: Summary of the candidate variants after filtering for variants observed in the commonly mutated genes list. A – Summary of the 17,286 variants in the KCH and SWiTCH filtered group. B – Summary of the 14,346 variants in the KCH, SWiTCH and TWiTCH group. C – Number of each variant type removed by filtering out Commonly Mutated Genes for both the KCH & SWiTCH (KS), and the KCH, SWiTCH & TWiTCH (KST) filtered groups.

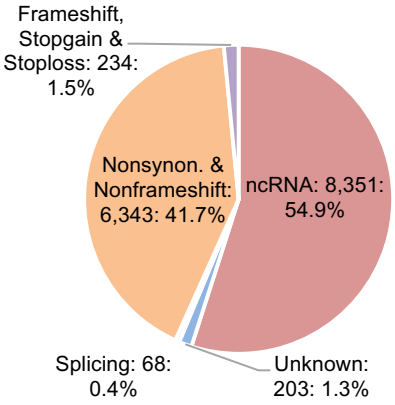
4.3.1.8 Haematopoiesis Associated Genes

This study aimed to identify variants that influence the severity of symptoms in SCA, as such, it was decided to narrow down the list of genes to those involved in erythroid development and function, as well as the immune response, which is known to play an important role in SCA disease pathology.

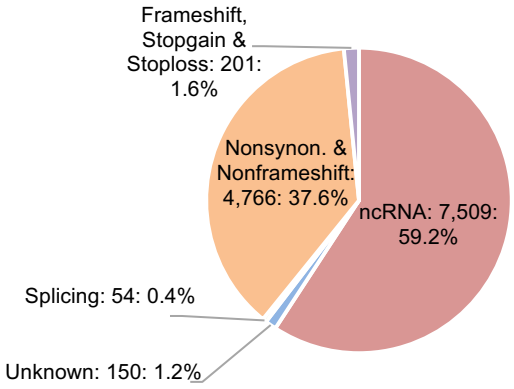
To achieve this, a list of 7,420 'haematopoietically silent' genes was compiled, based on data from the FANTOM5 Consortium[422]. The FANTOM5 Consortium have a large collection of RNAseq data, from many tissues and cell lines in both mice and humans. The human tissues that were used to compile the haematopoietically silent gene list are shown in Appendix 6, and the full list of 7,420 genes and transcripts is provided in Appendix 11. Genes were defined as transcriptionally silent if they had an RNA expression level of <1tpm (transcripts per million) in all 11 of the tissues analysed.

The remaining candidate variants after removal of the haematopoietically silent genes are summarised in Figure 4.10. 2,087 and 1,666 variants were removed from the KCH and SWiTCH filtered, and the KCH SWiTCH and TWiTCH filtered groups, respectively. As was observed after removal of commonly mutated genes, the proportion of ncRNA variants was also increased after removal of haematopoietically silent genes, making up 33.8% and 30.4% of variants removed, representing 52.0% and 56.3% of variants before the filtering step.

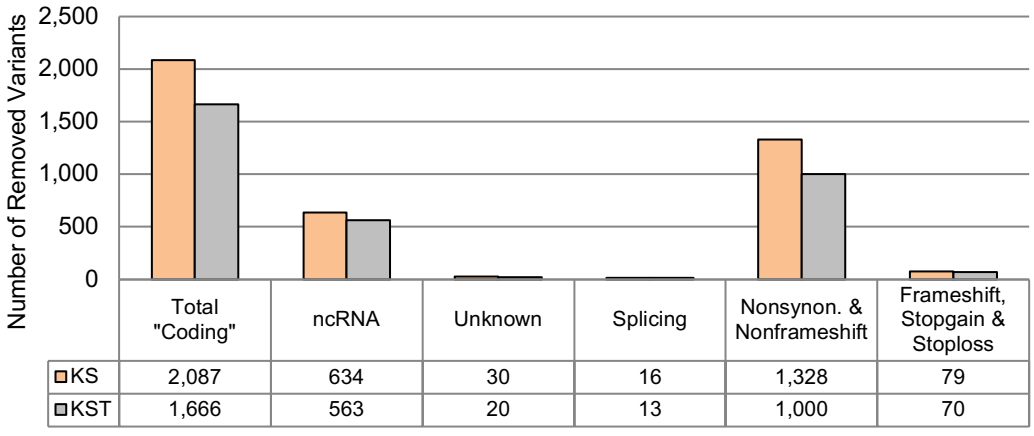
A



B



C



	Total "Coding"	ncRNA	Unknown	Splicing	Nonsynon. & Nonframeshift	Frameshift, Stopgain & Stoploss
KS	2,087	634	30	16	1,328	79
KST	1,666	563	20	13	1,000	70

Figure 4.10: Summary of the candidate variants after filtering for variants observed in the haematopoietically silent genes list. A – Summary of the 15,199 variants in the KCH and SWiTCH filtered group. B – Summary of the 12,680 variants in the KCH, SWiTCH and TWiTCH group. C – Number of each variant type removed by filtering out haematopoietically silent genes for both the KCH & SWiTCH (KS), and the KCH, SWiTCH & TWiTCH (KST) filtered groups.

4.3.1.9 Unknown Variants

Variants designated as 'UNKNOWN' by ANNOVAR are caused by errors in the transcript annotation and alignment process, and the 203 & 150 variants from the KCH and SWiTCH, and the KCH, SWiTCH and TWiTCH candidates respectively were excluded.

4.3.1.10 Allele Frequency – Rare Variants

In WES studies, it is common to filter for variants based on allele frequency in the general population, e.g. using data from the 1000 genomes project[423], with variants with a MAF >1% excluded. For studies aiming to identify variants causative of a disease phenotype, this is appropriate, since it is safe to assume that a mutation causing a disease will not be present in more than 1% of the population, unless 1% of the population are affected. This can be altered to 10% for recessive disorders, providing a homozygous genotype frequency of 1%, e.g. the β^S-globin SNP has a frequency of 3% in the 1000 genomes project data (10% in African populations), and so would be excluded by analyses looking for variants with a MAF <1%[423].

Since this study is aiming to identify modifiers of SCA, not the disease itself, it was decided that filtering based on allele frequency would not be appropriate, and any SNPs influencing disease severity may be relatively common in the rest of the population, with no observable phenotype. This is similar to what is observed in HPFH disorders, which are frequently asymptomatic and often only diagnosed in conjunction with a β-globinopathy.

4.3.1.11 CADD Phred-like Scores

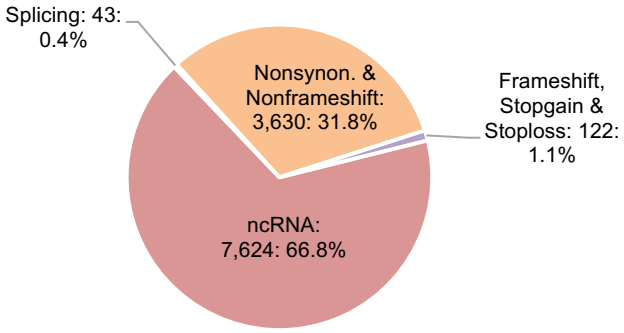
CADD (Combined Annotation-Dependent Depletion) scoring is a predictive measure of the likelihood of a variant to be deleterious to the function of the host gene, as modelled based on data from many different sources, including data regarding sequence conservation, GC content, CpG sites, transcription factor binding and other epigenetic data from the ENCODE project, transcript annotation and pre-existing amino acid change scoring tools[378]. The scores generated by CADD are converted to a Phred-like score, a scaled logarithmic ranking metric, where scores higher than 10 represent the top 10% most deleterious mutations, and higher than 20 represents the top 1%, 30 the top 0.1% etc. CADD Phred-like scoring was not used to filter variants in the initial analysis, but were used in conjunction with information about gene function to consider the plausibility for each of the top candidate variants in affecting SCA pathology.

4.3.1.12 Removal of Single Occurrence Genes

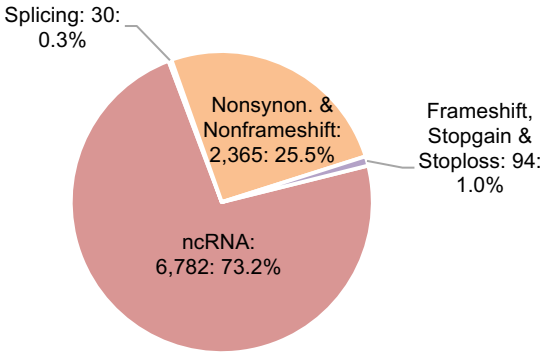
Any variants that occurred in only one patient were removed if the gene in which it was observed contained no other variants in the rest of the mild patients. This was because if a variant is not shared, nor the gene mutated in another mild patient, then an association cannot be shown. A summary of the two candidate groups is shown in Figure 4.11. 11,419 and 9,271 variants remain in the candidate gene list for the KCH and SWiTCH, and the KCH, SWiTCH and TWiTCH filtered groups, respectively, and are again enriched for ncRNA variants, which now represent more than two thirds of all remaining variants.

The candidate variant lists at this stage were used as the input for the gene burden test conducted in 4.4.

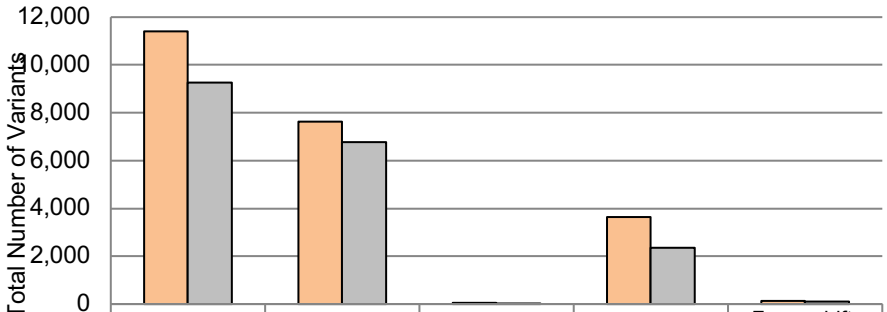
A



B



C



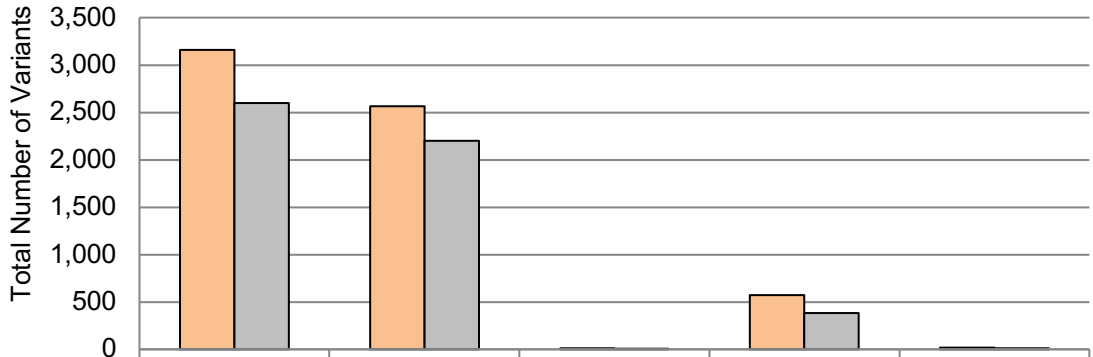
	Total "Coding"	ncRNA	Splicing	Nonsynon. & Nonframeshift	Frameshift, Stopgain & Stoploss
KS	11,419	7,624	43	3,630	122
KST	9,271	6,782	30	2,365	94

Figure 4.11: Summary of the candidate variants after exclusion of variants that occur only once, and in a gene that is not mutated in any other mild patient. A – Summary of the 11,419 variants in the KCH and SWiTCH (KS) filtered group. B – Summary of the 9,271 variants in the KCH, SWiTCH and TWiTCH group (KST). C – Comparison of each variant type for the KS and KST filtered groups.

4.3.2 Exome variants exclusive to mild patients

To identify individual variants with a high frequency in the mild patient group, variants that only occurred once in the mild patient group were excluded. Variants that were heterozygous in one of the severe groups could only be relevant under a recessive model, and so were excluded if not homozygous in at least two of the mild SCA patient group, even if heterozygous in multiple patients. The full list of 11,419 filtered variants is provided in Appendix 12.

The results of this final filtering step are shown in Figure 4.12, and demonstrate a significant reduction in the size of the candidate variant lists, suggesting that most of these variants were specific to one patient. The ncRNA still harbour the majority of variants, now representing 81.1% and 84.6% of all remaining variants.



	Total "Coding"	ncRNA	Splicing	Nonsynon. & Nonframeshift	Frameshift, Stopgain & Stoploss
KS	3,159	2,563	8	572	16
KST	2,597	2,197	6	382	12

Figure 4.12: Summary of the 3,159 and 2,597 candidate variants in the final lists for the KCH & SWiTCH (KS) and KCH, SWiTCH & TWiTCH (KST) filtered groups respectively. Loss of function variants (Splicing, Frameshift, Stopgain or Stoploss) were narrowed down to 24 and 18 variants in the KS and KST lists.

CADD Phred-like scores were obtained for candidate loss of function and coding mutations, and are displayed alongside the candidate variants in 4.3.2.1 & 4.3.2.2.

4.3.2.1 Nonsense: Loss of Function Candidate Variants

After the removal of variants that occurred only once in the dataset, there were eight remaining variants that affected splicing, ten that resulted in a frameshift, four that resulted in the gain of a stop codon, and two that resulted in the loss of a stop codon. All 24 of these variants are shown in Table 4.4, ranked by frequency within the mild group. The variants that were present in the TWiTCH severe group but absent from SWiTCH and KCH severe groups are also shown.

							Mild Group		Severe
Chr	*Pos*	*Var*	*Gene*	*Type*	*Details*	*CADD Phred score*	*Hom*	*Het*	*Het*
KCH, SWiTCH & TWiTCH Filtered Candidate Loss of Function Variants									
11	124972553	C:T	TMEM218	Stopgain	W3X	5.18	0	9	0
4	53611484	C:T	ERVMER34-1	Stopgain	W68X	35.00	0	5	0
7	99269397	T:C	CYP3A5	Stoploss	X141W	2.43	0	5	0
4	2721759	T:C	FAM193A	Stoploss	X1212R	5.47	0	4	0
15	101550877	A:-	LRRK1	Splicing	ex9 -2	26.30	2	1	0
7	44112998	C:-	POLM	Frameshift Del	R397fs	3.04	2	1	47
2	234474229	-:A	USP40	Frameshift Ins	L3fs	19.40	1	2	0
17	80755631	A:-	TBCD	Splicing	ex8 -2	26.00	1	2	0
11	704605	A:-	TMEM80	Frameshift Del	T271fs	0.41	0	3	0
12	133698497	A:-	ZNF891	Frameshift Del	V3fs	14.57	0	3	0
2	165657066	T:C	COBLL1	Splicing	ex4 -2	10.05	0	3	0
8	74169207	C:A	C8orf89	Splicing	ex3 +1	11.68	0	3	0
13	114059901	T:-	LOC101928841	Frameshift Del	T868fs	1.31	0	2	0
10	6010779	-:G	IL15RA	Frameshift Ins	C102fs	0.26	0	2	0
1	168105586	-:G	GPR161	Frameshift Ins	P17fs	10.29	0	2	0
19	17397478	-:TT	ANKLE1	Frameshift Ins	V637fs	9.29	0	2	0
3	46305948	T:G	CCR3	Splicing	ex2 +2	6.22	0	2	0
12	92382871	A:G	C12orf79	Splicing	ex4 +2	1.80	0	2	0
KCH & SWiTCH Filtered Candidate Loss of Function Variants – Not present in KST									
8	21966711	G:-	NUDT18	Frameshift Del	P35fs	13.01	5	0	0
19	39739155	T:-	IFNL4	Splicing	ex2 -2	15.66	2	2	0
17	74073456	CCGTCCTGGC:-	GALR2	Frameshift Del	P370fs	26.50	0	3	0
16	4414415	C:G	CORO7	Splicing	ex13 -1	25.30	0	2	0
16	4519398	G:A	NMRAL1	Stopgain	R37X	35.00	0	2	0
7	99758145	C:T	GAL3ST4	Stopgain	W289X	38.00	0	2	0

Table 4.4: 24 variants resulting in splice site disruption, frameshift, stoploss or stopgain in the mild SCA patient group after filtering. 6 of these were absent from the SWiTCH group but observed in the TWiTCH group. fs indicates frameshift, and splice variants are annotated as exN +/- 1/2, where the variant is either one or two nucleotides upstream (-) or downstream (+) of exon N.

One of the stopgain mutations shown in Table 4.4 occurs in NMRAL1, and is heterozygous in two mild patients. NMRAL1 detects changes in the NADPH/$NADP^+$ ratio, and regulates downstream signalling pathways, including NO synthesis, with signalling modulated through interaction with argininosuccinate synthetase, and also regulates apoptosis through the NF-κB pathway[424,425]. The variant in NMRAL1 substitutes arginine at position 37 for a stop codon, in an

exon included in all three annotated transcripts. This results in a truncated version of the 299 residue peptide, and the variant has a high CADD Phred-like score of 35.00.

The 24 loss of function variants occur in genes involved in a broad range of biological pathways and processes, some of which present a plausible biological mechanism to affect the pathophysiology of SCA. The most interesting candidate from this point of view is NMRAL1, which regulates synthesis of NO, a signalling molecule that has already been associated with the disease. This variant was predicted to effect gene function, as indicated by a high CADD Phred-like score, and is expected to completely remove functionality from all transcripts on the affected allele.

Loss of function variants affect the amount of gene product produced in the cell, and an important factor to consider when investigating these variants is whether or not the cell can rescue these levels by compensating with increased expression of other isoforms, or from other alleles. Many of the variants identified in this analysis occurred in a limited proportion of splicing variants, and depending on the individual roles of each these isoforms, and any redundancy in function shared between them, any effect on gene function in the cell may be mediated. Similarly, most of the variants shown in Table 4.4 are heterozygous, including the stopgain mutation in the candidate gene NMRAL1.

4.3.2.2 Missense: Nonsynonymous Substitutions and Nonframeshift Insertions/Deletions

After the removal of variants that occurred only once in the dataset, there were 509 nonsynonymous substitutions remaining, as well as 27 nonframeshift deletions and 35 nonframeshift insertions. Nonframeshift insertions or deletions are characterised by addition or removal of a sequence fragment consisting only of complete codons (i.e. with sequence length being a multiple of three), with the result that in the protein sequence individual amino acids are added or lost, but the downstream sequence does not change. These sequence changes are generally considered to be less deleterious to gene function than frameshift mutations, which change the reading frame of the downstream sequence, essentially resulting in a nonsense protein sequence that can result in the use of a premature translational termination site, or is disposed of at the mRNA level by nonsense mediated decay[426].

The 20 of the 572 nonsynonymous and nonframeshift candidate variants with the highest frequency in the mild SCA patient group are shown in Table 4.5.

							Mild Group		Severe
Chr	*Pos*	*Var*	*Gene*	*Type*	*Details*	*CADD Phred Score*	*Hom*	*Het*	*Het*
KCH, SWiTCH & TWiTCH Filtered Candidate Nonsynonymous & Non-Frameshift Variants									
19	7293898	G:C	INSR	Nonsyn. SNP	A2G	0.50	16	0	0
9	139222174	T:C	GPSM1	Nonsyn. SNP	V8A	0.25	15	0	0
12	132313109	-:TGCCGCTGC	MMP17	Non-FS Ins	P24LPLP	3.28	13	2	0
6	1313952	A:G	FOXQ1	Nonsyn. SNP	E338G	0.01	9	6	0
1	2126139	C:G	C1orf86	Nonsyn. SNP	R17P	13.23	13	1	0
13	28674628	T:C	FLT3	Nonsyn. SNP	D7G	16.24	7	7	0
20	60640315	AGGGCC:-	TAF4	Non-FS Del	183:184del	9.84	10	2	0
19	50155387	CGCTCC:-	SCAF1	Non-FS Del	581:582del	5.98	11	0	0
14	70039824	GGCGGC:-	CCDC177	Non-FS Del	171:172del	7.26	10	1	0
2	231902471	C:G	C2orf72	Nonsyn. SNP	A64G	0.01	4	7	0
21	40178042	A:G	ETS2	Nonsyn. SNP	R140G	11.38	4	7	0
21	40178043	G:C	ETS2	Nonsyn. SNP	R140T	8.96	4	7	0
6	24797815	A:T	C6orf229	Nonsyn. SNP	F172L	23.90	3	8	0
22	19137658	G:A	GSC2	Nonsyn. SNP	R47C	15.68	3	8	0
9	139972219	G:T	UAP1L1	Nonsyn. SNP	V79L	0.00	2	9	0
17	73512653	G:T	TSEN54	Nonsyn. SNP	E4D	0.06	5	5	0
4	107279482	T:A	GIMD1	Nonsyn. SNP	K171I	3.72	0	10	0
19	55525508	G:A	GP6	Nonsyn. SNP	T602M	21.50	0	10	0
KCH & SWiTCH Filtered Candidate Nonsynonymous & Non-Frameshift – Not present in KST									
2	217498310	-:CGCTGCTGC	IGFBP2	Non-FS Ins	L22PLLL	12.54	14	0	0
1	33430102	T:G	RNF19B	Nonsyn. SNP	Q62P	6.34	13	1	0

Table 4.5: Summary of the top 20 candidate variants from the Nonsynonymous and non-frameshift substitutions after filtering. Ranked by frequency in the mild SCA patient group. Table shows 18 variants after filtering by the KCH, SWiTCH and TWiTCH groups, and 2 that were present in TWiTCH but absent from the KCH and SWiTCH severe exome datasets.

A non-frameshift insertion of three amino acid residues was observed in IGFBP2. This variant in was homozygous in 14 of the mild patients and results in insertion of proline-leucine-leucine before a leucine residue at position 22, and has a higher CADD Phred-like score of 12.54. This variant occurs in the signalling peptide sequence and is likely lost during post-translational processing of the protein. IGFBP2 encodes insulin-like growth factor binding protein 2, a

signalling molecule that promotes cell proliferation and differentiation, and has been associated with cancer progression[427,428]. IGFBP2 promotes HSC survival and expansion in bone marrow, as well as in cell culture conditions, and one of its targets (insulin-like growth factor 1) is used for the *in vitro* expansion of erythroid progenitors described in Chapter 30[172,429,430].

Substitutions were also observed in FOXQ1 (homozygous in 9, heterozygous in 6) and FLT3 (homozygous in 7, heterozygous in 7). FOXQ1 is a transcription factor activated by TGF-β/Wnt signalling, and is associated with colorectal cancer progression and metastasis[431–433]. Expression of FOXQ1 has been shown to be lost during the γ-globin to β-globin switch, and it is thought to transcriptionally repress BCL11A, one of the master regulators of this switch[434,435]. This variant causes the substitution of glutamic acid at position 338 for glycine, and has a very low CADD Phred-like score of 0.01. FLT3 is a receptor tyrosine kinase that stimulates the expansion of HSCs and erythroid progenitors in the bone marrow, and is commonly associated with blood cancers[436–438]. The variant in FLT3 results in an aspartic acid to glycine substitution at position 7, situated in the signal peptide. The substitutions in FLT3 has a CADD Phred-like score of 16.24, and is included in both isoforms of the protein.

A non-frameshift deletion of two amino acids was observed in TAF4, and which occurred in 12 mild SCA patients (10 homozygous and 2 heterozygous). TAF4 is a component of the transcription initiation complex, and interacts with the transcription factor CREB, which is thought to be capable of γ-globin re-activation[439,440]. The deletion in TAF4 results in the loss of a glycine-proline repeat at positions 183-184, with a CADD Phred-like score of 9.84.

Two variants were present adjacent to each other in ETS2, causing substitution of arginine at position 140 for either glycine or threonine. ETS2 is a transcription factor required for the formation of cardiac progenitor cells from fibroblasts during development, and is involved in the triggering of angiogenesis in endothelial cells[441,442]. Overexpression of ETS2 in K562 cells has also been shown to affect several important erythroid genes, reducing expression levels of KLF1, β-globin and α-globin[443]. Upon further investigation it was discovered that the two variants co-localised, and were observed in the same patients (four homozygous and seven heterozygous each). The fact that these were identified as two separate variants is due to mis-annotation, and rather than being two separate SNPs resulting in substitution of arginine to either glycine or threonine, this is a dinucleotide variant causing an arginine to alanine substitution. The CADD Phred-like scores for glycine and threonine substitutions were 11.38 and 8.96 respectively, and for the real substitution of alanine, was 10.57.

							Mild Group		Severe
Chr	*Pos*	*Var*	*Gene*	*Type*	*Details*	*CADD Phred Score*	*Hom*	*Het*	*Het*
KCH, SWiTCH & TWiTCH CADD Filtered Candidate Nonsynonymous & Non-Frameshift Variants									
1	2126139	C:G	C1orf86	Nonsyn. SNP	R17P	13.23	13	1	0
13	28674628	T:C	FLT3	Nonsyn. SNP	D7G	16.24	7	7	0
21	40178042	A:G	ETS2	Nonsyn. SNP	R140G	11.38	4	7	0
6	24797815	A:T	C6orf229	Nonsyn. SNP	F172L	23.90	3	8	0
22	19137658	G:A	GSC2	Nonsyn. SNP	R47C	15.68	3	8	0
19	55525508	G:A	GP6	Nonsyn. SNP	T602M	21.50	0	10	0
11	94261280	G:A	C11orf97	Nonsyn. SNP	R94K	16.30	2	7	0
9	131580998	C:T	ENDOG	Nonsyn. SNP	S12L	16.34	0	9	0
10	102770278	T:C	PDZD7	Nonsyn. SNP	K790E	12.36	1	6	0
X	100749038	C:T	ARMCX4	Nonsyn. SNP	A1821V	22.70	2	4	0
12	29936626	C:A	TMTC1	Nonsyn. SNP	R20L	12.65	1	5	0
11	22881002	C:T	CCDC179	Nonsyn. SNP	R29Q	10.45	5	6	4
3	113052314	G:C	CFAP44	Nonsyn. SNP	P1185R	26.90	1	4	0
12	55968284	T:C	OR2AP1	Nonsyn. SNP	L29P	24.60	1	4	0
4	107279518	T:C	GIMD1	Nonsyn. SNP	E159G	11.45	0	5	0
KCH & SWiTCH CADD Filtered Candidate Nonsynonymous & Non-Frameshift – Not present in KST									
2	217498310	-:CGCTGCTGC	IGFBP2	Non-FS Ins	L22PLLL	12.54	14	0	0
16	2077090	G:C	SLC9A3R2	Nonsyn. SNP	E28D	24.90	0	7	0
11	46369267	G:A	DGKZ	Nonsyn. SNP	A20T	23.40	0	6	0
19	1000785	C:T	GRIN3B	Nonsyn. SNP	H117Y	19.19	1	4	0
X	153707209	C:T	LAGE3	Nonsyn. SNP	D16N	10.78	1	4	0

Table 4.6: Summary of the top 20 candidate variants from the Nonsynonymous and non-frameshift substitutions after filtering, as in Table 4.5, with additional filtering of variants with CADD Phred-like scores <10. Variants are ranked by frequency in the mild SCA patient group. Table shows 15 variants after filtering by the KCH, SWiTCH & TWiTCH groups, and 5 that were present in TWiTCH but absent from the KCH and SWiTCH severe exome datasets. Seven candidate variants from Table 4.5 passed the CADD Phred-like score filtering.

Some of these candidate variants occur in genes with biologically plausible mechanisms to affect the SCA pathophysiology, but have very low CADD Phred-like scores, and are unlikely to affect gene function e.g. the variant in FOXQ1, which has been associated with repression of BCL11A, but has CADD Phred-like score of 0.01. To avoid this, it was decided to apply a CADD Phred-like score cut-off of 10.00 to the list of top candidate variants, to ensure that only those

predicted to affect gene function would be considered, the results of this are shown in Table 4.6. CADD Phred-like scores of >10.00 represent the top 10% of CADD scored variants, and are the most likely to be deleterious to gene function.

After the exclusion of candidate variants with CADD Phred-like scores of <10 from Table 4.5, only the variants in C1orf86, FLT3, ETS2, C6orf229, GSC2, GP6 and IGFBP2 were retained. Table 4.6 shows the updated candidate variant table after filtering by CADD Phred-like score, containing the new list of top candidates.

Of the variants identified by the analyses performed in this section, five candidates with biologically plausible mechanisms were observed. These are the variants in IGFBP2, FOXQ1, FLT3, TAF4 and ETS2. However, the variants in FOXQ1 and TAF4 had CADD Phred-like scores below the threshold of 10.00, and so were not predicted to be deleterious to gene function.

4.3.2.3 Non Protein Coding: ncRNA Candidate Variants

After the removal of variants that occurred only once in the dataset, there were 2,563 ncRNA variants remaining. The 20 ncRNA candidate variants with the highest frequency in the mild SCA patient group after filtering are shown in Table 4.7. It is worth noting that while CADD scoring is available for ncRNA variants, these scores are not included in Table 4.7, and were not used for assessment of ncRNA variants. Due to a the relative lack of annotated ncRNA data and functional understanding, the accuracy of CADD scoring is currently limited for ncRNA variants[378].

					Mild		Severe
Chr	*POS*	*Var*	*Gene*	*Type*	*Hom*	*Het*	*Het*
KCH, SWiTCH & TWiTCH Filtered ncRNA Variants							
8	29779220	T:C	FAM183CP	ncRNA exonic	15	0	0
20	61667631	T:C	LINC00029	ncRNA exonic	10	4	0
20	25658484	A:G	ZNF337-AS1	ncRNA exonic	12	1	0
17	68063631	G:A	LINC01028	ncRNA exonic	11	2	0
8	11618998	G:C	C8orf49	ncRNA exonic	12	2	1
8	73663426	T:A	LOC101926908	ncRNA exonic	5	7	0
19	38042410	C:G	ZNF571-AS1	ncRNA exonic	2	10	0
1	149576483	G:A	LINC00623,LINC00869,LOC103091866	ncRNA exonic	2	10	0
3	156392191	T:C	TIPARP-AS1	ncRNA exonic	11	0	0
10	17428971	T:A	ST8SIA6-AS1	ncRNA exonic	10	1	0
14	95999998	G:C	SNHG10	ncRNA exonic	4	7	0
12	104260287	T:C	GNN	ncRNA exonic	3	8	0
10	91597426	T:C	LINC00865	ncRNA exonic	3	8	0
20	25834293	C:T	LOC101926935	ncRNA exonic	1	10	0
11	65272383	C:T	MALAT1	ncRNA exonic	1	10	0
13	22849324	C:T	LINC00540	ncRNA exonic	10	0	0
2	67313607	T:C	LOC102800447	ncRNA exonic	8	2	0
7	1201192	A:T	LOC101927021	ncRNA exonic	5	5	0
KCH & SWiTCH Filtered ncRNA Variants – Not present in KST							
13	114452024	A:G	LINC00552	ncRNA exonic	9	3	0
19	51398377	G:A	KLKP1	ncRNA exonic	7	5	0

Table 4.7: Summary of the top 20 candidate variants from the ncRNA candidate variants after filtering. Variants are ranked by frequency in the mild SCA patient group. Table shows 18 variants after filtering by the KCH, SWiTCH & TWiTCH groups, and two that were present in TWiTCH but absent from the KCH & SWiTCH severe exome datasets.

Of the 20 ncRNA variants shown in Table 4.7, 19 have no characterised function, this demonstrates the vast gap in our current knowledge about the individual function of these non-coding transcripts.

MALAT1 is the only ncRNA in Table 4.7 with a characterised function, and presents a plausible biological mechanism to influence the severity of SCA. MALAT1 is a long ncRNA that regulates serine/arginine splicing factors and is associated with cancer progression, negatively regulating the tumour suppressor gene p53, as well as B-MYB[444,445]. B-MYB is a key regulator of HSC proliferation, and when depleted results in a reduction of haematopoietic potential. Through this interaction, MALAT1 regulates early haematopoietic development[446,447].

All of these variants are exonic SNPs, which are unlikely to affect transcriptional interference, and without more information regarding the function and mechanism of action of these ncRNA, it is difficult to predict how their function would be affected. SNPs could affect the formation of secondary structures required for the ncRNA function, and any ncRNA that work by sequence

specific binding could be affected, but may also have a degree of tolerance for short mismatches[448].

4.3.3 Variants in Known Modifier Genes

Genes that had previously been associated with the SCA phenotype were screened for variants in the mild SCA patient population. This was performed using the list of 14,996 KCH and SWiTCH filtered candidate variants generated after removal of 'unknown' variants in 4.3.1. These include genes from both the α-globin-like and β-globin-like gene loci, as well as some known regulators of globin locus expression (MYB, BCL11A, KLF1, ASH1L, GATA1, LMO2, and LDB1). The results of this search are shown in Table 4.8.

The only variant observed in one of the β-globin-like genes occurred in β-globin (HBB), and was confirmed to correspond to the β^0 thalassaemia mutation in patient SCD 215. It is reassuring that this variant made it through the filtering process, and also that it has a very high CADD Phred-like score of 33.00, since it is known to have a severe effect on gene function.

							Mild Group	
Chr	*Pos*	*Var*	*Gene*	*Type*	*Details*	*CADD Phred score*	*Hom*	*Het*
1	155327167	A:T	ASH1L	Nonsyn. SNP	I2332N	12.82	0	1
1	155491102	T:C	ASH1L	Nonsyn. SNP	Q70R	24.60	0	1
1	155449342	T:C	ASH1L	Nonsyn. SNP	I1107V	0.63	0	1
1	155531805	T:A	ASH1L-AS1	ncRNA	n/a	13.62	0	1
1	155533060	G:T	ASH1L-AS1	ncRNA	n/a	3.86	0	1
1	155532525	C:G	ASH1L-AS1	ncRNA	n/a	7.98	0	1
1	155532696	-:C	ASH1L-AS1	ncRNA	n/a	5.19	0	3
11	5247979	-:T	HBB	Frameshift Ins	D48fs	33.00	0	1
16	230574	T:C	HBQ1	Nonsyn. SNP	L30P	24.70	0	1
19	12995802	T:C	KLF1	Nonsyn. SNP	H329R	27.50	0	1

Table 4.8: Results of a search for variants in the candidate gene list that occur in known modifier genes for SCA phenotype severity. 10 variants were identified, all of which were heterozygous. One variant occurs in the β-globin gene (HBB) in patient SCD 215, who was heterozygous for both HbS and β^0 thalassaemia, as described in Table 4.2. This frameshift variant is the β^0 mutation, since it prevents any functional β-globin expression from this allele.

There was also only one variant observed in the α-globin–like genes, which occurred in θ-globin (HBQ1). This variant resulted in a single amino acid substitution of leucine to proline at position 30, and also had a high CADD Phred-like score of 24.70. θ-globin is an α-globin homologue that is highly conserved, but is only transcribed at very low levels[449,450]. Given the protective effect of

persistent ζ-globin expression in SCA, it is feasible that any SNP that increases expression of θ-globin could also ameliorate the severity of the SCA phenotype.

A SNP was observed in KLF1, substituting histidine for arginine at position 329 in the second of three zinc finger domains, with a very high CADD Phred-like score of 27.50. As described in 1.6.1, variants affecting the function of KLF1 could impair the γ-globin to β-globin switch during erythroid development. This variant was observed in patient SCD 213, which had the highest HbF level observed in the mild patient group, of 29.5%.

Three SNPs were also identified in the histone methyltransferase ASH1L, with CADD Phred-like scores of 24.60, 12.82 and 0.63, and a further four variants were observed in ASH1L-AS1, an ncRNA running antisense to the ASH1L gene.

4.4 Analysis 1: Gene Burden Analysis

It is possible that individual variants protective for the SCA severe phenotype are not shared between the patients in the SCA mild group, but occur in the same genes. To investigate this model, a simple gene burden test was performed.

This was carried out using the list of 11,419 candidate variants generated after filtering of genes with one variant in one patient as described in 4.3.1.12, rather than the list of 3,159 generated after removal of single occurrence variants generated in 4.3.2, since many variants that occur only once in the Mild SCA patient group would still be relevant if additional variants are observed in that gene in other patients.

For each candidate variant, a list of the mild SCA patients in which that SNP is observed was generated, these were then collated for each gene, giving information on how many of the 19 patients contained a candidate variant in each gene, and how many different variants were observed in that gene. Results were ranked firstly by the number of the mild SCA patients affected, and secondly by the fewest number of variants. Genes with fewer variants were prioritised since it was expected that the most variable genes have a higher tolerance for sequence polymorphisms, as is the case for the 'Commonly Mutated Gene' list described in 4.3.1.7.

The top 20 candidate genes from this analysis are shown in Table 4.9, firstly showing the list inclusive of ncRNA, which are shown to be highly variable, with many mutated in most of the patients, and with many variants. An additional candidate gene list was generated with the ncRNAs removed. Candidate genes that contained only one variant were also excluded, since individual variants with a high frequency in the mild SCA patient group had already been investigated by the analyses in 4.3.2.

The majority of genes identified by the gene burden test were observed in ncRNA, this is probably at least partially due to the fact that ncRNA made up the majority of the candidate variant list, accounting for 66.8% of the 11,419 variants analysed from the KCH and SWiTCH filtered group in Figure 4.11. The results in Table 4.9 also show that the ncRNA generally harbour far more variants than the coding genes, and the average number of variants per ncRNA is 19.2, compared to 6.7 in the coding genes. This is to be expected given that ncRNA function is generally more tolerant of individual sequence variation[448]. However there are some exceptions to this, for example, the ncRNA LOC100287944 has only two variants that affect 18

of the 19 patients, whereas the MUC22 coding gene has 31 variants. MUC22 is a member of the mucin gene family, the majority of which were filtered out by the 'Commonly Mutated Gene' list described in 4.3.1.7.

ncRNA Included			
Gene Name	*Number of Patients*	*Number of Variants*	*Type of Gene*
MRGPRG-AS1	19	6	ncRNA
SLC6A10P	19	6	ncRNA
FAM215A	19	7	ncRNA
LINC00955	19	19	ncRNA
GUCY2EP	19	24	ncRNA
MUC22	19	31	Coding
LOC101926935	19	47	ncRNA
LOC401357	19	51	ncRNA
MST1L	18	2	Coding
LOC100287944	18	2	ncRNA
LOC100133077	18	7	ncRNA
ESPNP	18	8	ncRNA
LINC00469	18	9	ncRNA
LINC00940	18	12	ncRNA
LINC01262	18	13	ncRNA
LINC00552	18	13	ncRNA
LOC100129697	18	14	Coding
LINC00937	18	14	ncRNA
KCNQ1OT1	18	43	ncRNA
PRKXP1	18	45	ncRNA

ncRNA Removed		
Gene Name	*Number of Patients*	*Number of Variants*
MUC22	19	31
MST1L	18	2
LOC100129697	18	14
TUBGCP3	17	3
MMP17	16	2
C11orf97	16	2
IGFBP2	16	3
C1orf86	16	3
DSPP	16	11
LOC100129520	16	12
SUCLG2	15	3
FOXQ1	15	4
BEAN1	14	2
RNF225	14	2
OR1D5	14	3
GIMD1	14	4
ARMCX4	14	7
RALGDS	14	7
GOLGA8H	14	9
STARD9	14	10

Table 4.9: Top 20 candidate genes identified by the gene burden test, ranked by the number of mild patients containing a variant in each gene, and by the total number of variants observed in the gene. Table on the left shows the list including ncRNA, and on the right shows only protein coding genes.

Of the 17 ncRNA shown in Table 4.9, only one has any characterised function. KCNQ1OT1 is a chromatin interacting ncRNA, it binds to the KCNQ1 locus and recruits chromatin modifying complexes, promoting formation of heterochromatin and transcriptionally silencing the surrounding genes[451,452]. Dysregulation of at the KCNQ1 locus has previously been associated with Beckwith-Wiedemann syndrome and Silver Russell Syndrome[453,454].

Of the 20 protein coding genes, seven have no known function (MST1L, LOC100129697, C11orf97, LOC100129520, RNF225, ARMCX4 and GOLGA8H). Variants in MMP17, IGFBP2 and C1orf86 were previously identified in the analysis shown in 4.3.2.2, with individual variants accounting for 15, 14 and 14 of the mild patients respectively. A variant in FOXQ1 was

previously identified in 4.3.2.2, with this variant found in 15 patients, three of these patients also harbour an extra variant in the FOXQ1 gene.

Of the top candidate genes identified by the gene burden test, the only gene that presents a biologically plausible mechanism to affect the SCA disease pathophysiology is the BCL11A repressor FOXQ1. However, the most frequent variant in this gene (E338G) was investigated in 4.3.2.2, and was not predicted to affect gene function, with a very low CADD Phred-like score of 0.01. Excluding this SNP, the remaining three FOXQ1 variants are each heterozygous in one patient.

4.5 Fisher's Exact Tests

Statistical analyses were used to test for associations of individual variants with the SCA patient groups outlined in 4.2. These were performed independently of the variant filtering pipeline, so as to not introduce inherent biases based on the filtering criteria.

After the removal of low quality variant calls, as defined by the ANNOVAR variant calling pipeline, the number of occurrences of each variant within each patient group was counted. These counts were used to test the statistical significance of the frequency of each variant in these groups, using three separate Fisher's exact tests:

1. Firstly, a simple patient count test was performed, testing for the number of patients carrying each variant within the group, disregarding whether they are homozygous or heterozygous.
2. The second test was for allele frequency, considering each patient as two alleles with 0, 1 or 2 copies of the variant. This investigates whether there is an imbalance in the distribution of these alleles between the investigated groups.
3. The third test was for homozygous counts. This was to account for a model where the effect of the variant is recessive, and only the homozygous patients would be affected. In this case the heterozygous cases in either group were ignored.

4.5.1 P-Values & Multiple Testing Correction

For most scientific experiments, a p-value cut off of <0.05 is used to test for statistical significance. This threshold signifies a 5% probability of the observed difference occurring purely by chance, resulting in a false positive rejection of the null hypothesis. Therefore, if performing a test on twenty different variables, it would be expected that at least one of them would incorrectly reject the null hypothesis, regardless of whether there is a true effect or not. For studies where multiple variables are investigated, it is common practice to adjust the p-value cut-off to account for this multiple testing. The Bonferroni correction is a simple conversion of this value, where the desired p-value threshold is divided by the number of tests, e.g. if testing 20 variables with a p-value cut off of 0.05, the Bonferroni correction would be 0.05/20, giving a p-value threshold of <0.0025 [455]. For studies investigating genome-wide datasets, every base pair sequenced could potentially be considered an additional variable, although this is generally considered to be prohibitively conservative, and it is common practice

for genome-wide studies to Bonferroni correct using the number of common SNPs covered by the sequencing area (with common SNPs defined as having a MAF of >5%). An estimate of this for whole genome sequencing is 4,152,114, and for exome sequencing using the Agilent SureSelect capture kit is 58,091, resulting in adjusted p-value thresholds of $p<1.2x10^{8}$ and $p<8.6x10^{7}$ respectively[456]. For the Roche NimbleGen capture kit, this estimation is 50,000 common variants, resulting in a cut-off of $p<1.0x10^{6}$ [456].

Table 4.10 shows the number of common variants identified in our own datasets, for both the Agilent SureSelect and Roche NimbleGen kits respectively, compared to those estimated by Lacey *et al.*[456].

Group	*Total Variants*	*Common Variants*	*P-Value*	*Capture Kit*	*Estimated Common Variants*	*Estimated P-Value*
Mild	2,798,560	2,180,113	$2.29x10^{8}$	Agilent SureSelect	58,091	$8.61x10^{7}$
Severe	566,010	449,433	$1.11x10^{7}$	Agilent SureSelect	58,091	$8.61x10^{7}$
SWiTCH	1,712,023	1,089,120	$4.59x10^{8}$	Roche NimbleGen	50,000	$1.00x10^{6}$
TWiTCH	2,036,852	1,378,096	$3.63x10^{8}$	Roche NimbleGen	50,000	$1.00x10^{6}$
HUSTLE	1,966,311	1,310,841	$3.81x10^{8}$	Roche NimbleGen	50,000	$1.00x10^{6}$
Unknown	3,808,700	2,829,713	$1.77x10^{8}$	Roche NimbleGen	50,000	$1.00x10^{6}$

Table 4.10: Table summarising the number of common variants (minor allele frequency >5% in the 1000 Genomes Project data[191]) for each of the SCA patient groups, and comparing to those estimated for the same exome capture kits by Lacey *et al.* [456]. The numbers of common variants are much higher than expected, resulting in much stricter Bonferroni corrected p-value thresholds.

The total number of common variants annotated from both the Roche and Agilent exome capture kits is drastically higher than suggested by Lacey *et al.* and subsequently the corresponding p-value thresholds are lower[456,457]. Due to this discrepancy, rather than using these estimates for the Bonferroni correction for each exome capture kit, a new correction factor will be calculated for each analysis performed, corresponding to the number of variants being tested.

4.5.2 Analysis 2: Statistical Comparison of Mild & Severe SCA Patient Groups

To identify variants associated with either the mild or severe SCA patient phenotype, variants from the group of 19 mild SCA patients were compared to those in the severe group, firstly including just the five KCH severe patients, and secondly including the data from the 132 SWiTCH patients as well.

4.5.2.1 Mild and severe SCA patients from King's College Hospital

The initial analyses performed considered only the patients from King's College Hospital, the ten most significant results from the 2,922,494 variants tested are shown for each of the three Fisher's Exact Tests in Table 4.11, Table 4.12 & Table 4.13. The difference in p-values between the different tests demonstrate that the most significant results were identified by the allele frequency test, although this is likely due to the fact that each patient is considered as two separate alleles, effectively doubling the sample number.

						Mild		*Severe*	
Chr	*Position*	*Var*	*Gene*	*Type*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
3	153725251	T:C	C3orf79, ARHGEF26-AS1	Intergenic	2.35E-05	16	3	0	0
16	13359272	C:T	SHISA9, ERCC4	Intergenic	2.35E-05	14	5	0	0
9	66489205	A:G	LINC01410, PTGER4P2-CDK2AP2P2	Intergenic	2.35E-05	4	15	0	0
21	10935167	TG:-	TPTE	Intronic	2.35E-05	0	19	0	0
4	190544967	A:G	LINC01060, LINC01262	Intergenic	0.000141	1	0	5	0
15	90172474	A:G	KIF7	Intronic	0.000141	1	0	5	0
10	42400172	A:T	LOC441666	Intergenic	0.000141	0	1	0	5
10	70139112	T:G	RUFY2	Intronic	0.000141	0	1	0	5
8	43094814	C:G	HGSNAT, POTEA	Intergenic	0.000141	0	1	0	5
8	43094823	C:T	HGSNAT, POTEA	Intergenic	0.000141	0	1	0	5

Table 4.11: Patient count test. The ten variants with the lowest p-values as tested by Fisher's Exact Test for patient counts between the mild and severe SCA groups. Analysis includes all variants annotated in patients from King's College London only. The lowest p-value is 2.35×10^{-5}, and does not reach the threshold of 1.71×10^{-8} required for statistical significance after Bonferroni Correction. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the mild or severe patient groups.

None of the variants in the mild or severe patient groups reached statistical significance when tested for significance by the patient count test in Table 4.11, or the homozygous count test in Table 4.13.

However, statistical significance was reached using the allele frequency test in Table 4.12, finding variants significantly enriched in both the severe group compared to the mild, and vice versa. The ten most significant variants from the allele frequency test are shown in Table 4.12, and seven of these pass the Bonferroni corrected p-value threshold of $<1.71 \times 10^{-8}$.

						Mild		Severe	
Chr	*Position*	*Var*	*Gene*	*Type*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
4	190544967	A:G	LINC01060, LINC01262	Intergenic	1.01E-08	1	0	5	0
15	90172474	A:G	KIF7	Intronic	1.01E-08	1	0	5	0
1	246677104	A:C	SMYD3, LOC255654	Intergenic	1.01E-08	18	0	0	0
2	54885314	A:T	SPTBN1	Intronic	1.01E-08	18	0	0	0
2	64881017	C:G	SERTAD2	UTR5	1.01E-08	18	0	0	0
2	64881018	G:C	SERTAD2	UTR5	1.01E-08	18	0	0	0
2	71148045	C:T	VAX2	Intronic	1.01E-08	18	0	0	0
12	63073959	G:T	PPM1H	Intronic	4.37E-08	17	1	0	0
2	74856123	C:T	M1AP	Intronic	4.37E-08	17	1	0	0
7	81358712	G:A	HGF	Intronic	4.37E-08	17	1	0	0

Table 4.12: Allele frequency test. The ten variants with the lowest p-values as tested by Fisher's Exact Test for allele frequency between the mild and severe SCA groups. Analysis includes all variants annotated in patients from King's College London only. Only p-values for the first seven variants fall below the Bonferroni corrected threshold of 1.71×10^{-8} for statistical significance. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the mild or severe patient groups.

Comparing the most significant results of the allele frequency test to those of the homozygous count test (Table 4.12 & Table 4.13 respectively), it is clear that there is a lot of overlap. This is expected given that the variants with the highest allele frequency will often be those that are homozygous in the most patients. Interestingly those variants that reach statistical significance in the allele frequency test are far above the threshold in the homozygous count test, presumably due to the doubling of the sample number in the former. This demonstrates how the technique used in the analyses can artificially alter the significance of the findings.

						Mild		Severe	
Chr	*Position*	*Var*	*Gene*	*Type*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
4	190544967	A:G	LINC01060, LINC01262	Intergenic	0.000141	1	0	5	0
15	90172474	A:G	KIF7	Intronic	0.000141	1	0	5	0
1	246677104	A:C	SMYD3, LOC255654	Intergenic	0.000141	18	0	0	0
2	54885314	A:T	SPTBN1	Intronic	0.000141	18	0	0	0
2	64881017	C:G	SERTAD2	UTR5	0.000141	18	0	0	0
2	64881018	G:C	SERTAD2	UTR5	0.000141	18	0	0	0
2	71148045	C:T	VAX2	Intronic	0.000141	18	0	0	0
2	92315552	T:A	ACTR3BP2	Intergenic	0.000471	0	0	4	0
10	125622253	A:-	CPXM2	Intronic	0.000471	19	0	1	3
12	126517773	G:A	LINC00939, LOC101927464	Intergenic	0.000471	0	11	4	1

Table 4.13: Homozygous count test. The ten variants with the lowest p-values as tested by Fisher's Exact Test for homozygous patient count between the mild and severe SCA groups. Analysis includes all variants annotated in patients from King's College London only. The lowest p-value is 0.000141, and does not reach the threshold of 1.71×10^{-8} required for statistical significance after Bonferroni Correction. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the mild or severe patient groups.

4.5.2.2 Mild and Severe including SWiTCH Trial Exomes

Patient count, allele frequency and homozygous count tests for variants enriched in either the mild or severe SCA patient groups were repeated, with the SWiTCH trial patients included in the severe group. The top ten most significant of the 3,860,685 variants tested are shown for each these tests in Table 4.14, Table 4.15 and Table 4.16.

Seven of the ten most significant variants from the patient count test in Table 4.14 are intergenic.

						Mild		*Severe*	
Chr	*Position*	*Var*	*Gene*	*Type*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
16	13359272	C:T	SHISA9, ERCC4	Intergenic	8.16E-25	14	5	0	0
12	48418556	G:T	COL2A1, SENP1	Intergenic	1.63E-23	15	4	1	0
5	137024472	C:T	KLHL3	Intronic	1.63E-23	15	4	1	0
2	91694185	T:C	LOC654342	Intergenic	1.63E-23	1	18	1	0
3	153725251	T:C	C3orf79, ARHGEF26-AS1	Intergenic	1.63E-23	16	3	1	0
6	31336926	-:TT	HLA-B, MICA	Intergenic	1.63E-23	15	4	0	1
4	45615916	T:G	GNPDA2, GABRG1	Intergenic	1.63E-23	17	2	1	0
11	71140156	C:T	FLJ42102, DHCR7	Intergenic	1.63E-23	9	10	1	0
14	57116979	A:G	TMEM260	Downstream	1.63E-23	18	1	1	0
3	96712868	A:T	EPHA6	Intronic	1.63E-23	6	13	0	1

Table 4.14: Patient count test. 10 most significant variants from Fisher's Exact Test for patient count between Mild and Severe groups, including 132 severe patients from SWiTCH. The lowest p value is 8.16×10^{-25}, and all ten of these variants reach the significance threshold of 1.30×10^{-8}. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the mild or severe patient groups.

						Mild		*Severe*	
Chr	*Position*	*Var*	*Gene*	*Type*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
18	8336695	-:GAAGGG	PTPRM	Intronic	7.13E-47	19	0	1	0
13	42017682	T:G	OR7E37P	ncRNA	1.02E-44	19	0	2	0
19	48564545	A:G	PLA2G4C	Intronic	1.02E-44	19	0	2	0
4	8790898	G:A	CPZ, HMX1	Intergenic	1.02E-44	19	0	2	0
14	57116979	A:G	TMEM260	Downstream	1.85E-44	18	1	1	0
12	63073959	G:T	PPM1H	Intronic	3.2E-43	17	1	0	0
12	125670964	A:C	AACS, TMEM132B	Intergenic	6.45E-43	19	0	3	0
12	125670970	A:C	AACS, TMEM132B	Intergenic	6.45E-43	19	0	3	0
15	91590948	A:T	LOC101926911, SV2B	Intergenic	6.45E-43	19	0	3	0
15	95111778	A:G	MCTP2, LOC440311	Intergenic	6.45E-43	19	0	3	0

Table 4.15: Allele Frequency Test. 10 most significant variants from Fisher's Exact Test for allele frequency between Mild and Severe groups, including 132 severe patients from SWiTCH. The lowest p value is 7.13×10^{-47}, and all ten of these variants reach the significance threshold of 1.30×10^{-8}. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the mild or severe patient groups.

Table 4.15 shows the most significant variants from the allele frequency test, one SNP, the downstream mutation in TMEM260 was also present in the patient count and homozygous count tests in Table 4.14 and Table 4.16.

						Mild		*Severe*	
Chr	*Position*	*Var*	*Gene*	*Type*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
18	8336695	-:GAAGGG	PTPRM	Intronic	1.63E-23	19	0	1	0
13	42017682	T:G	OR7E37P	ncRNA	1.71E-22	19	0	2	0
19	48564545	A:G	PLA2G4C	Intronic	1.71E-22	19	0	2	0
4	8790898	G:A	CPZ, HMX1	Intergenic	1.71E-22	19	0	2	0
12	125670964	A:C	AACS, TMEM132B	Intergenic	1.26E-21	19	0	3	0
12	125670970	A:C	AACS, TMEM132B	Intergenic	1.26E-21	19	0	3	0
12	2039690	A:G	LINC00940	ncRNA	1.26E-21	19	0	3	2
15	91590948	A:T	LOC101926911, SV2B	Intergenic	1.26E-21	19	0	3	0
15	95111778	A:G	MCTP2, LOC440311	Intergenic	1.26E-21	19	0	3	0
17	33307703	C:-	LIG3	Intronic	1.26E-21	19	0	3	1

Table 4.16: Homozygous count test. 10 most significant variants from Fisher's Exact Test for homozygous patient count between Mild and Severe groups, including 132 severe patients from SWiTCH. The lowest p value is 1.63×10^{-23}, and all ten of these variants reach the significance threshold of 1.30×10^{-8}. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the mild or severe patient groups.

The ten most significant results from the homozygous count test are shown in Table 4.16. As was observed previously in 4.5.2.1, there was a large overlap between the most significant variants from the allele frequency test and the homozygous count test. Only two of the variants in Table 4.16 were absent from Table 4.15.

These results present interesting, statistically significant candidate variants located at gene loci that could have a biologically plausible effect on the SCA phenotype, e.g. SMYD3, SENP1, KLHL3 and TMEM123B. However, in these cases it seems very unlikely that any of the variants are able to influence the gene function, being mostly intergenic or intronic, and not disrupting annotated transcription factor binding sites. In contrast, some variants have a plausible mechanism for disruption of gene regulation, e.g. being situated in a densely populated transcription factor binding site, or deletion of CG dinucleotides from a CpG island, but affect genes with functions that seem very unlikely to influence the SCA phenotype.

The ten most significant variants for each of the three analyses all reach the Bonferroni corrected p-value threshold required for statistical significance. This shows that including the exome data for the US SCA patients from the SWiTCH clinical trial greatly increases the power to identify statistically significant differences between the two groups. However, due to the size

of the SWiTCH dataset, and the fact that it is made up of SCA patients recruited from the US, the two groups being compared are no longer separated solely by SCA phenotype severity, but also by geographic location and ethnic ancestry.

Of the variants presented in these results, none occur in the coding region. Intergenic, intronic and other non-coding regions have a much higher tolerance for genetic variation, since they are less likely to have a serious impact on gene function, and therefore are not strongly selected for or against. This genetic variation therefore accumulates in the non-coding areas of the genome over time, and can produce a lot of 'noise' when performing genome-wide analyses. It is worth noting that this is not always the case, and that there are some important non-coding regulatory elements associated with SCA and the Beta globin locus e.g. the 'Corfu' deletion in an intergenic region at the β-globin locus, that results in dysregulation and increased HBF expression (described in more detail in 1.3)[458].

4.5.2.3 Mild and Severe including SWiTCH, with non-coding variants removed

In order to avoid the noise generated by the large number of non-coding variants, the analyses performed in 4.5.2.2 were repeated, but this time with the intergenic, intronic, upstream and downstream mutations removed.

As described in 2.2.1.1 and 2.2.1.2, two different exome capture kits were used for library preparation. Agilent SureSelect was used for all 19 mild patients, as well as the 5 severe patients from King's College Hospital, while Roche NimbleGen was used for the US dataset that makes up the rest of the severe patient group. The coverage of the genome varies between different capture kits, especially in the non-coding regions, and could therefore identify false-positive associations purely due to the sampling technique used for each group.

To account for any discrepancies in the ncRNA targeted, lists of ncRNA were generated for both the SureSelect and NimbleGen groups, and each ncRNA was only included in the study if at least one variant in one sample was observed in both the SureSelect and the NimbleGen groups. 4988 ncRNA were found to be shared between the groups, out of a total of 5324 and 5778 for SureSelect and NimbleGen datasets respectively, the full list is shown in Appendix 7, and is summarised in 4.3.1.4.

The results of the analyses are shown in Table 4.17, Table 4.18 and Table 4.19, for the patient count test, allele frequency test, and homozygous count test respectively. A full list of the 2,442 significant variants is included in Appendix 8.

							Mild		Severe	
Chr	*Position*	*Var*	*Gene*	*Type*	*CADD Phred score*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
7	77326410	T:C	APTR	ncRNA	6.86	1.71E-22	15	4	2	0
13	42017682	T:G	OR7E37P	ncRNA	0.06	1.71E-22	19	0	2	0
X	6975782	C:G	HDHD1	Exonic SNP	1.77	1.26E-21	10	9	1	2
18	77440128	T:G	CTDP1	Exonic SNP	0.00	2.13E-21	12	6	1	0
5	150311858	T:-	ZNF300P1	ncRNA splicing	8.19	7.23E-21	10	9	4	0
1	31973125	GAGTCT GTCTG:-	LINC01225	ncRNA	9.51	7.23E-21	7	12	1	3
1	31973409	G:A	LINC01225	ncRNA	7.15	7.23E-21	7	12	1	3
21	15646397	A:G	ABCC13	ncRNA	2.64	7.23E-21	12	7	4	0
5	150311622	A:G	ZNF300P1	ncRNA	1.68	7.23E-21	12	7	4	0
5	150311678	A:G	ZNF300P1	ncRNA	0.11	7.23E-21	12	7	4	0

Table 4.17: Filtered patient count test. 10 most significant variants from Fisher's Exact Test for patient count between Mild and Severe groups, including 132 severe patients from SWiTCH. The lowest p value is 1.71×10^{-22}, and all ten of these variants reach the significance threshold of 2.29×10^{-8}. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the mild or severe patient groups. Intergenic, intronic, downstream and upstream variants have been removed, along with ncRNA exclusive to one exome capture kit.

After the removal of the non-coding variants, variants that affect amino acid sequence are observed among the most significant results, such as the SNP in HDHD1, a pseudouridine-5'-phosphatase that is active in erythrocytes, and is involved in the processing of by-products of RNA degradation[459].

							Mild		Severe	
Chr	*Position*	*Var*	*Gene*	*Type*	*CADD Phred Score*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
13	42017682	T:G	OR7E37P	ncRNA	0.06	1.02E-44	19	0	2	0
12	2039690	A:G	LINC00940	ncRNA	0.57	2.38E-41	19	0	3	2
2	186655726	G:A	FSIP2	Exonic SNP	6.58	2.38E-41	19	0	3	2
7	128294446	A:G	LINC01000	ncRNA	2.16	2.38E-41	19	0	4	0
17	41961451	T:C	MPP2	Splicing	0.70	1.25E-40	19	0	4	1
4	156706482	G:A	GUCY1B3	Splicing	2.79	1.25E-40	19	0	4	1
X	65382685	T:C	HEPH	Exonic SNP	0.00	1.25E-40	19	0	4	1
3	149376058	T:G	WWTR1-AS1	ncRNA	5.76	3.08E-40	18	0	2	0
3	51990119	A:C	GPR62	Exonic SNP	5.20	3.08E-40	18	0	2	0
1	221507141	C:T	C1orf140	ncRNA	5.12	5.98E-40	19	0	5	0

Table 4.18: Filtered allele frequency test. 10 most significant variants from Fisher's Exact Test for allele frequency between Mild and Severe groups, including 132 severe patients from SWiTCH. The lowest p value is 1.02×10^{-44}, and all ten of these variants reach the significance threshold of 2.29×10^{-8}. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the mild or severe patient groups. Intergenic, intronic, downstream and upstream variants have been removed, along with ncRNA exclusive to one exome capture kit.

The most significant results from the filtered allele frequency and homozygous count tests are shown in Table 4.18 and Table 4.19. The ten variants identified are the same for the two tests, differing only in the order in which the significance is ranked e.g. MPP2 has a lower ranking in

the allele frequency test than the homozygous count test, since it is observed in a heterozygous severe patient as well as the four homozygous severe patients, whereas this is ignored in the homozygous count test.

							Mild		Severe	
Chr	*Position*	*Var*	*Gene*	*Type*	*CADD Phred score*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
13	42017682	T:G	OR7E37P	ncRNA	0.06	1.71E-22	19	0	2	0
12	2039690	A:G	LINC00940	ncRNA	0.57	1.26E-21	19	0	3	2
2	186655726	G:A	FSIP2	Exonic SNP	6.58	1.26E-21	19	0	3	2
17	41961451	T:C	MPP2	Splicing	0.70	7.23E-21	19	0	4	1
4	156706482	G:A	GUCY1B3	Splicing	2.79	7.23E-21	19	0	4	1
7	128294446	A:G	LINC01000	ncRNA	2.16	7.23E-21	19	0	4	0
X	65382685	T:C	HEPH	Exonic SNP	0.00	7.23E-21	19	0	4	1
3	149376058	T:G	WWTR1-AS1	ncRNA	5.76	2.11E-20	18	0	2	0
3	51990119	A:C	GPR62	Exonic SNP	5.20	2.11E-20	18	0	2	0
1	221507141	C:T	C1orf140	ncRNA	5.12	3.47E-20	19	0	5	0

Table 4.19: Filtered homozygous count test. 10 most significant variants from Fisher's Exact Test for homozygous patient count between Mild and Severe groups, including 132 severe patients from SWiTCH. The lowest p value is 1.71×10^{-22}, and all ten of these variants reach the significance threshold of 2.29×10^{-8}. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the mild or severe patient groups. Intergenic, intronic, downstream and upstream variants have been removed, along with ncRNA exclusive to one exome capture kit.

A coding variant was observed in HEPH. HEPH encodes Hephaestin, a ferroxidase involved in the processing of iron in intestinal cells, and delivery into the blood[460]. This provides an obvious biological mechanism by which sickle cell severity could be affected, lower blood iron levels would limit haemoglobin production, perhaps to below the threshold required to aggregate and distort the erythrocyte membrane. This would alleviate the symptoms caused by vaso-occlusive events, but severe anaemia would still be observed.

Two splicing variants were identified in MPP2 and GUCY1B3, both of which are homozygous in all 19 mild patients and four severe patients, as well as being heterozygous in one severe patient. GUCY1B3 is a subunit of guanylate cyclase, which acts as the receptor for the NO signalling pathway, it directly recognises NO and is then activated to convert GTP to cGMP, triggering the downstream signalling cascade[461]. Mutations at this locus have previously been associated with hypertension and cardiovascular disease[462]. Impaired guanylate cyclase could either reduce response to NO signalling, or even leave it constitutively active. NO is a vasodilator that has previously been linked to the pathophysiology of SCA and pulmonary hypertension, and is thought to contribute to the mechanism of action of HU[24,241,244,463].

Some of these variants occur in genes that have a plausible biological mechanism for influencing SCA phenotype, including HDHD1, HEPH and GUCY1B3. However, many others

can be excluded based on their expression patterns, these include FSIP2, which is expressed exclusively in spermatocytes, and is therefore unlikely to affect SCA disease pathophysiology.

Table 4.17, Table 4.18 and Table 4.19 also include the CADD Phred-like score for each of the variants identified. None of the variants have a CADD Phred-like score higher than 10.00, and so do not fall within the top 10% of variants predicted to affect gene function. The variants in HDHD1, HEPH and GUCY1B3 have particularly low CADD Phred-like scores of 1.77, 0.001 and 2.79 respectively, and are not predicted to affect gene function.

4.5.2.4 Most of the significant variants associate with ancestry, not disease severity

Many of the variants highlighted as being statistically significant between the groups are likely to be due to the difference in ancestry of the UK SCA patients compared to the USA SCA patients. In this study, all 19 mild patients were recruited from SCA patients living in London, whereas the severe group consists of 132 patients recruited from the USA, plus an additional 5 from London. Of the variants shown in Table 4.17, Table 4.18 and Table 4.19, the SNPs are present in either 18 or 19 of the 19 mild patients, and no more than 5 of the severe patients. This sort of imbalance is to be expected given that the tests are looking to identify imbalances between the two groups. However, upon further investigation it was discovered that apart from the variant in CTDP1, the candidate variants present in the severe group were only observed in the five patients that were recruited from the UK. In the case of CTDP1, the homozygous severe case was from the US dataset.

For example, the ncRNA SNP in C1orf140 was statistically significant due to the fact that it is homozygous in all 19 patients from the mild group, and is homozygous in only 5 out of the 137 severe patients. However this can also be viewed as being homozygous in all 24 of the UK SCA patients, and completely absent from the 132 US patients. It is unlikely that this is due to the difference in the capture kits used between the two datasets, since the ncRNA filtering step that was applied ensures that only ncRNA with variants present in the data from both capture kits are included.

4.5.3 Analysis 3: Statistical Comparison of SWiTCH and HUSTLE SCA Patient Groups

The analyses described previously focus on the identification of variants either protective or causative of the severe SCA phenotype, by identifying imbalances in frequency in our mild SCA patient group from KCH compared to a severe group made up of mostly US patients.

Taking a different approach, it was decided to investigate variants enriched in either the SWiTCH or HUSTLE patient groups within the US dataset. As described in 4.2.2, the SWiTCH group contains severe SCA patients, having had a stroke before the age of 17.5, whereas the HUSTLE group shows no particular bias for severity, and any SCA patient receiving HU therapy at St Jude's Children's Research Hospital could be included. Based on these assumptions, a comparison of these groups could inform on any variants enriched (or depleted) in the stroke group compared to the general SCA population.

4.5.3.1 SWiTCH and HUSTLE Fisher's Exact Tests

Three Fisher's Exact Tests were performed on the variants identified in SWiTCH and HUSTLE exome groups, testing for patient count, allele frequency and homozygous count, as was performed for the mild and severe groups described in 4.5.2. The results of these analyses are shown in Table 4.20, Table 4.21 and Table 4.22 respectively, and it can be seen that the majority of the most significant variants are non-coding, similar to the observations in the mild vs severe patient tests in 4.5.2.1 and 4.5.2.2. A total of 2,673,201 variants were tested, giving a Bonferroni corrected p-value threshold for statistical significance of 1.87×10^{-8}.

						SWiTCH		*HUSTLE*	
Chr	*Position*	*Var*	*Gene*	*Type*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
16	12009279	A:G	GSPT1	Exonic SNP	8.26E-17	91	4	30	1
4	151177432	A:C	DCLK2	UTR3	4.35E-16	60	21	17	3
2	166810373	A:G	TTC21B	Upstream	4.35E-16	81	0	20	0
11	209002	A:G	RIC8A	Intronic	4.99E-16	75	15	25	3
5	101596078	TATAT:-	SLCO4C1	Intronic	6.59E-16	51	24	12	4
14	88852283	G:-	SPATA7	Intronic	7.95E-16	96	0	34	0
9	33264540	C:G	BAG1	Exonic SNP	8.08E-16	69	10	17	2
21	38081577	C:G	SIM2	Intronic	8.08E-16	61	18	12	7
19	56041255	C:G	SBK2	Exonic SNP	8.08E-16	78	1	19	0
1	2518186	A:G	FAM213B	Upstream	8.77E-16	86	0	25	0

Table 4.20: Patient count test. The ten variants with the lowest p-values as tested by Fisher's Exact Test for patient counts between the SWiTCH and HUSTLE SCA groups. The lowest p-value is 8.26×10^{-17}, and all ten variants reach the threshold of 1.87×10^{-8} required for statistical significance after Bonferroni Correction. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the SWiTCH or HUSTLE patient groups.

Three protein coding variants were identified in Table 4.20, including a SNP in BAG1. BAG1 is a cell cycle regulator that upregulates anti-apoptotic factors, including BCL2, and has been shown to be essential for healthy development of both haematopoietic and neuronal cells in mice[464,465].

SWiTCH *HUSTLE*

						SWITCH		HUSTLE	
Chr	*Position*	*Var*	*Gene*	*Type*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
16	12009279	A:G	GSPT1	Exonic SNP	4.37E-31	91	4	30	1
2	166810373	A:G	TTC21B	Upstream	4.66E-31	81	0	20	0
14	88852283	G:-	SPATA7	Intronic	1.65E-30	96	0	34	0
1	2518186	A:G	FAM213B	Upstream	3.39E-30	86	0	25	0
19	56041255	C:G	SBK2	Exonic SNP	4.43E-30	78	1	19	0
1	1229680	G:T	ACAP3	Intronic	5.61E-30	101	0	40	0
1	115631924	C:T	TSPAN2	Intronic	1.51E-29	74	5	16	4
22	31687142	G:C	PIK3IP1	Intronic	2.16E-29	79	0	20	0
1	246670281	G:C	SMYD3	Intronic	6.96E-29	81	0	22	0
5	149737221	T:G	TCOF1	UTR5	7.55E-29	77	0	19	0

Table 4.21: Allele frequency test. The ten variants with the lowest p-values as tested by Fisher's Exact Test for allele frequency between the SWiTCH and HUSTLE SCA groups. The lowest p-value is 4.37×10^{-31}, and all ten variants reach the threshold of 1.87×10^{-8} required for statistical significance after Bonferroni Correction. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the SWiTCH or HUSTLE patient groups.

						SWITCH		HUSTLE	
Chr	*Position*	*Var*	*Gene*	*Type*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
2	166810373	A:G	TTC21B	Upstream	4.35E-16	81	0	20	0
14	88852283	G:-	SPATA7	Intronic	7.95E-16	96	0	34	0
1	2518186	A:G	FAM213B	Upstream	8.77E-16	86	0	25	0
1	1229680	G:T	ACAP3	Intronic	1.11E-15	101	0	40	0
1	115631924	C:T	TSPAN2	Intronic	1.69E-15	74	5	16	4
16	12009279	A:G	GSPT1	Exonic SNP	1.69E-15	91	4	30	1
19	56041255	C:G	SBK2	Exonic SNP	2.08E-15	78	1	19	0
22	31687142	G:C	PIK3IP1	Intronic	2.91E-15	79	0	20	0
5	149737221	T:G	TCOF1	UTR5	5.30E-15	77	0	19	0
1	246670281	G:C	SMYD3	Intronic	5.36E-15	81	0	22	0

Table 4.22: Homozygous count test. The ten variants with the lowest p-values as tested by Fisher's Exact Test for homozygous patients between the SWiTCH and HUSTLE SCA groups. The lowest p-value is 4.35×10^{-16}, and all ten variants reach the threshold of 1.87×10^{-8} required for statistical significance after Bonferroni Correction. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the SWiTCH or HUSTLE patient groups.

As was observed in the mild vs severe Fisher's Exact Tests performed in 4.5.2.1 and 4.5.2.2, the majority of the most significant variants occur in the non-coding region. In these previous tests no coding variants were identified, however in the SWiTCH vs HUSTLE tests, three exonic SNPs were observed, in GSPT1, SBK2, and BAG1, which has been linked to healthy regulation of haematopoietic cells[465].

4.5.3.2 SWiTCH and HUSTLE Fisher's Exact Test with Non-Coding Variants Removed

Table 4.23, Table 4.24 and Table 4.25 show the most significant results from the three Fisher's Exact Tests shown in 4.5.3.1, with the non-coding variants removed. CADD Phred-like scores are also included in these tables. A full list of the 236 significant variants is included in Appendix 9. Three of these coding variants (in GSPT1, BAG1 and SBK2) were described in the previous section. The variants observed in GSPT1 and SBK2 had very low CADD Phred-like scores of

0.001 and 0.02 respectively, and are not predicted to affect gene function. The variant in BAG1 however had a high CADD Phred-like score of 16.29.

								SWiTCH		*HUSTLE*	
Chr	*Position*	*Var*	*Gene*	*Type*	*Details*	*CADD Phred score*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
16	12009279	A:G	GSPT1	Nonsyn. SNP	V100A	0.00	8.26E-17	91	4	30	1
9	33264540	C:G	BAG1	Nonsyn. SNP	G45R	16.29	8.08E-16	69	10	17	2
19	56041255	C:G	SBK2	Nonsyn. SNP	A298P	0.02	8.08E-16	78	1	19	0
19	10676681	T:C	KRI1	Nonsyn. SNP	T5A	0.00	3.99E-15	54	26	15	6
7	6193521	G:C	USP42	Nonsyn. SNP	R779P	11.14	5.30E-15	56	21	12	7
5	140537363	C:T	PCDHB17P	ncRNA	n/a	19.17	1.01E-14	76	3	20	1
19	51015404	T:C	ASPDH	Nonsyn. SNP	Q266R	9.60	2.19E-14	75	10	24	3
1	1361641	C:T	TMEM88B	Nonsyn. SNP	P45L	5.81	2.66E-14	62	36	32	8
19	4670313	C:G	MYDGF	Nonsyn. SNP	G12R	22.90	2.74E-14	31	37	4	10
5	54830295	T:G	RNF138P1	ncRNA	n/a	0.53	3.29E-14	51	24	16	3

Table 4.23: Patient count test. The ten variants with the lowest p-values as tested by Fisher's Exact Test for patient counts between the SWiTCH and HUSTLE SCA groups, with non-coding variants removed. The lowest p-value is 8.26×10^{-17}, and all ten variants reach the threshold of 1.87×10^{-8} required for statistical significance after Bonferroni Correction. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the SWiTCH or HUSTLE patient groups.

A variant with a high CADD Phred-like score of 22.90 was observed in MYDGF, resulting in the substitution of glycine to arginine at position 12 in the signal peptide. MYDGF is a myeloid derived growth factor, produced by monocytes and macrophages in response to ischaemic tissue damage in the heart, preventing apoptosis by activating the PI3K signalling pathway, as well as stimulating endothelial cell growth and angiogenesis in the damaged tissue[466]. MYDGF has been proposed as a therapeutic treatment to repair cardiac tissues after ischaemic injuries, and if its ability to minimise tissue damage and reperfusion injury are observed in other tissues as well, MYDGF could have an important role in modulating the severity of SCA symptoms triggered by vaso-occlusive events[466].

								SWiTCH		HUSTLE	
Chr	*Position*	*Var*	*Gene*	*Type*	*Details*	*CADD Phred score*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
16	12009279	A:G	GSPT1	Nonsyn. SNP	V100A	0.00	4.37E-31	91	4	30	1
19	56041255	C:G	SBK2	Nonsyn. SNP	A298P	0.02	4.43E-30	78	1	19	0
5	140537363	C:T	PCDHB17P	ncRNA	n/a	19.17	1.21E-27	76	3	20	1
9	33264540	C:G	BAG1	Nonsyn. SNP	G45R	16.29	2.75E-27	69	10	17	2
5	132149684	G:C	SOWAHA	Nonsyn. SNP	R124P	9.51	2.42E-26	76	1	21	0
8	120220779	G:-	MAL2	Frameshift Del	V23fs	24.90	3.56E-26	91	0	34	0
7	140396475	-:G	NDUFB2-AS1	ncRNA	n/a	8.17	4.20E-26	94	0	37	0
16	25704145	A:G	HS3ST4	Nonsyn. SNP	Q136R	7.82	6.80E-26	65	7	14	3
1	151881885	A:C	THEM4	Nonsyn. SNP	L17R	0.09	9.36E-26	62	5	12	2
8	145106943	CC:-	OPLAH	Frameshift Del	R1166fs	23.90	1.07E-25	73	0	19	0

Table 4.24: Allele frequency test. The ten variants with the lowest p-values as tested by Fisher's Exact Test for allele frequency between the SWiTCH and HUSTLE SCA groups, with non-coding variants removed. The lowest p-value is 4.31×10^{-31}, and all ten variants reach the threshold of 1.87×10^{-8} required for statistical significance after Bonferroni Correction. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the SWiTCH or HUSTLE patient groups.

								SWiTCH		HUSTLE	
Chr	*Position*	*Var*	*Gene*	*Type*	*Details*	*CADD Phred score*	*P Value*	*Hom*	*Het*	*Hom*	*Het*
16	12009279	A:G	GSPT1	Nonsyn. SNP	V100A	0.00	1.70E-15	91	4	30	1
19	56041255	C:G	SBK2	Nonsyn. SNP	A298P	0.02	2.08E-15	78	1	19	0
5	140537363	C:T	PCDHB17P	ncRNA	n/a	19.17	4.56E-14	76	3	20	1
8	120220779	G:-	MAL2	Frameshift Del	V23fs	24.90	1.12E-13	91	0	34	0
7	140396475	-:G	NDUFB2-AS1	ncRNA	n/a	8.17	1.27E-13	94	0	37	0
5	132149684	G:C	SOWAHA	Nonsyn. SNP	R124P	9.51	1.5E-13	76	1	21	0
8	145106943	CC:-	OPLAH	Frameshift Del	R1166fs	23.9	1.94E-13	73	0	19	0
4	48492434	G:C	ZAR1	Nonsyn. SNP	Q42H	0.01	2.3E-13	70	3	17	2
16	25704145	A:G	HS3ST4	Nonsyn. SNP	Q136R	7.82	3.82E-13	65	7	14	3
9	33264540	C:G	BAG1	Nonsyn. SNP	G45R	16.29	5.45E-13	69	10	17	2

Table 4.25: Homozygous count test. The ten variants with the lowest p-values as tested by Fisher's Exact Test for homozygous patient counts between the SWiTCH and HUSTLE SCA groups, with non-coding variants removed. The lowest p-value is 1.70×10^{-15}, and all ten variants reach the threshold of 1.87×10^{-8} required for statistical significance after Bonferroni Correction. Hom and Het refer to the number of patients found to be homozygous or heterozygous respectively, within the SWiTCH or HUSTLE patient groups.

Table 4.25 shows the most significant variants identified by the homozygous count test. Nine of the variants shown were also identified in the allele frequency test in Table 4.24, likely due to the fact that the allele frequency test favours homozygous variants.

Of the variants identified by these analyses, two SNPs in BAG1 and MYDGF present plausible biological mechanisms for influencing the severity of the SCA phenotype. Both of these variants have CADD Phred-like scores >10.00 (16.29 and 22.90 respectively), and are significantly enriched in the SWiTCH severe patient group compared to the HUSTLE group. The G45R variant in BAG1 was among the most significant variants for all three tests, being present in 79

of 132 SWiTCH patients (69 homozygous and ten heterozygous), and 19 of 140 HUSTLE patients (17 homozygous and two heterozygous). The G12R variant in MYDGF was observed in 68 of the 132 SWiTCH patients (31 homozygous and 37 heterozygous), and 14 of the 140 HUSTLE patients (four homozygous and ten heterozygous), and was only included among the most significant variants for the patient count test, not the allele frequency of homozygous count tests, since the majority of the patients were heterozygous. This MYDGF variant had a p-value of 1.80×10^{-19} for the allele frequency test, reaching the threshold of $p < 1.87 \times 10^{-8}$ required for significance. However for the homozygous count test the p-value was 2.5×10^{-7}, and was not significant, suggesting that if this SNP is affecting the SCA phenotype, it is acting under a dominant model, and that homozygosity is not required.

4.6 Summary of the SCA WES Results

Using WES data generated by this study and in combination with publicly available datasets, we were able to identify nine potential modifiers of the SCA disease phenotype (Table 4.26). These nine candidates occur in genes with biologically plausible mechanisms to influence the pathophysiology of the disease, and warrant further testing *in vitro* through the use of CRISPR genomic editing.

							Mild		*Severe*
Chr	*Pos*	*Var*	*Gene*	*Type*	*Details*	*CADD Phred score*	*Hom*	*Het*	*Het*
Analysis 1: Variants in known modifier genes									
16	230574	T:C	HBQ1	Nonsyn. SNP	L30P	24.7	0	1	0
19	12995802	T:C	KLF1	Nonsyn. SNP	H329R	27.5	0	1	0
Analysis 1: Loss of function variants									
16	4519398	G:A	NMRAL1	Stopgain	R37X	35.00	0	2	0
Analysis 1: Missense variants									
2	217498310	-:CGCT GCTGC	IGFBP2	Non-FS Ins	L22PLLL	12.54	14	0	0
13	28674628	T:C	FLT3	Nonsyn. SNP	D7G	16.24	7	7	0
2	40178042	AG:GC	ETS2	Nonsyn. SNP	R140A	10.57	4	7	0
Analysis 1: ncRNA variants									
11	65272383	C:T	MALAT1	ncRNA exonic	n/a	n/a	1	10	0

							SWiTCH		*HUSTLE*	
Chr	*Pos*	*Var*	*Gene*	*Type*	*Details*	*CADD Phred score*	*Hom*	*Het*	*Hom*	*Het*
Analysis 3: Variants enriched in SWiTCH or HUSTLE patients										
9	33264540	C:G	BAG1	Nonsyn. SNP	G45R	16.29	69	10	17	2
19	4670313	C:G	MYDGF	Nonsyn. SNP	G12R	22.9	31	37	4	10

Table 4.26: Table summarising the nine candidate modifier variants identified by the different exome sequencing analysis strategies used. 7 of these variants were identified in the mild SCA patient group from KCH using the variant filtering pipeline developed in Analysis 1. Two variants in BAG1 and MYDGF were identified by Fisher's Exact Tests for enrichment in either the SWiTCH or HUSTLE SCA exome groups.

Seven of these modifier variants were identified in the Mild SCA patient group and are predicted to protect patients from the severe SCA phenotype, whilst two were identified as being enriched in the SWiTCH cohort relative to the HUSTLE cohort, and are predicted to increase phenotypic severity.

The two variants in KLF1 & HBQ1 were identified by searching for variants in known modifier genes that had made it through the variant filtering pathway, and both would be expected to elicit their effect on the SCA phenotype by de-repressing early stage globin genes (γ-globin and θ-globin respectively).

Five other candidate variants were identified by Analysis 1, and are present in more than one of the mild SCA patients at KCH, and absent from the severe SCA groups from KCH and SWiTCH, having made it through the variant filtering pipeline. These five variants include a loss of function mutation in NMRAL1, three missense mutations in IGFBP2, FLT3 and ETS2, and one variant in the ncRNA MALAT1. The variant in NMRAL1 would be expected to affect nitric oxide signalling, while IGFBP2, FLT3, ETS2 and MALAT1 have all previously been associated with haematopoietic regulation.

Two candidate variants in BAG1 and MYDGF were identified by Analysis 3, being significantly enriched in the severe SWiTCH cohort compared to the HUSTLE cohort, which is representative of the general SCA population. Therefore, these variants are predicted to increase rather than ameliorate the severity of the SCA phenotype. The variant in BAG1 is expected to affect haematopoietic regulation, while MYDGF has previously been associated with recovery from ischaemic injury.

The potential mechanisms by which these mutations may affect the SCA phenotype, and how they could be functionally analysed in future work, are discussed in more detail in 6.2.4.

If the predicted impact of these candidates on gene function is confirmed *in vitro*, this will demonstrate the power of WES and our variant filtering pipeline as a tool for identification of phenotype altering variants, especially given the small sample size of the mild SCA patient group. This would also support our hypothesis that many candidate genetic modifiers of SCA remain to be identified, and would demonstrate the importance of carrying out genome-wide sequencing studies on a larger scale to identify additional genetic factors influencing phenotype severity in global SCA populations.

Chapter 5 Results: CRISPR Genomic Editing - Functional Analysis of SNPs *in vitro*

WES provides a powerful tool for identification of variants associated with disease phenotypes. However, after *in silico* identification, functional analyses must be conducted to determine whether candidate variants are actually causative of the phenotype, or whether the observed association is coincidental. The advent of CRISPR genomic editing has greatly improved our ability to introduce specific mutations into the genome, and is a powerful tool for analysing the effect that specific variants have on gene function *in vitro*.

We plan to use CRISPR genomic editing to functionally assess the candidate variants that were identified by the WES study performed in Chapter 4. Prior to identification of these variants, we set out to familiarise ourselves with the CRISPR-Cas9 system, and to set up a pipeline in our laboratory to allow efficient introduction of candidate variants into cell lines *in vitro*.

This part of the thesis work aimed to test and optimise the CRISPR pipeline by investigating two previously identified variants that are thought to affect regulation of the β-globin locus. These two variants affect ASH1L and KLF1, and their discovery and proposed mechanisms of action are described below. We aimed to introduce these variants into a K562 cell line, and to perform preliminary functional analyses to determine what effect the mutations have on globin gene expression. K562 is an erythroleukaemic cell line, and is commonly used as a model for the erythroid lineages. The reasons for selecting K562 cells is described in more detail below, in 5.2.3.

Cas9 plasmids containing gRNAs designed to target the regions of interest were constructed as described in 2.4.2, initially using DNA repair templates incorporated into the Cas9 plasmids. As is described below in 5.1, this technique was chosen for simplicity, to allow transfection of a single plasmid that contained all three of the components for our CRISPR-Cas9 system.

5.1 CRISPR-Cas9 Strategy and Design

As described in 1.7.3, there are now a wide variety of CRISPR-based techniques and methodologies routinely used in laboratories around the world, that can be tailored and implemented according to the specific requirements of the intended experiments. While this makes CRISPR-Cas9 a powerful and versatile tool for laboratory research, the wealth of

information available also makes it very complex when initially establishing a CRISPR based protocol in a laboratory, due to the numerous options available.

5.1.1 gRNA & Template Sequence Design

When targeting the introduction of specific mutations, the options for gRNA design are quite limited, since the distance between the DSB and the target is a key factor for determining successful incorporation rates of the mutation[467]. gRNA design therefore involved a compromise between distance from the SNP site, computationally predicted cleavage efficiency, and number of predicted off-target binding sites (described in detail in 2.4.1).

Design of the template sequence is obviously restricted regarding the target SNP, but is less specific regarding disruption of the PAM site. Disruption of the PAM site prevents repeated cleavage of the target sequence once the template is correctly incorporated, and so increases efficiency. However, introducing additional sequence changes may also affect gene function independently of the SNP of interest. Since the KLF1 SNP is intronic, and hypothesised to affect transcription factor binding, it is therefore possible that the PAM disruption mutation could also affect the transcription factor binding at that site. In the case of the ASH1L SNP, the sequence allows the disruption of the PAM site without altering the amino acid sequence of the protein, but could potentially have an impact on protein levels as a result of altered codon usage. For the chosen PAM disruptions, asparagine AAC to AAT have usage frequencies of 0.54 and 0.46 respectively, while arginine CGG to CGA have 0.21 and 0.11[468].

It was decided that the benefits of increased efficiency outweighed these risks, and PAM disruption mutations were included in the template sequence. Additional PAM disruption only controls were also designed, in order to assay the effects that the PAM disruption mutations themselves have on gene function.

5.1.2 Delivery Methods for gRNA, Cas9 & Template Sequence

A straightforward approach was initially chosen for this project, where all three components of the system (Cas9, gRNA and template DNA) were included in a single plasmid (Figure 2.1). Plasmid transfection of K562 cells is well established, and this allowed the use of a single GFP reporter to confirm that all the necessary components had been successfully introduced into the cell. The use of a plasmid for the source of each of the components also makes the technique more cost-effective, since once constructed, the plasmid can be cloned and modified with

minimal additional cost in the laboratory, rather than purchasing each of the individual components for each experiment.

The main disadvantage of this technique was that it required a relatively large plasmid (approximately 10kb), which had adverse effects on transfection efficiency, (as described in 5.3.1). While it can be seen that this did not have a significant impact on cleavage efficiency in cells that had been successfully transfected, it was believed that the low copy number of the template DNA had an adverse effect on HDR rates. In response to this, the protocol was subsequently modified, and co-transfection with siRNA to knock down key components of the NHEJ pathway (Ku70 and Ligase IV), as well as separate introduction of the template sequence in the form of ssODNs were tested. Each of these techniques had previously been demonstrated to increase template uptake[353,469].

Template sequences for insertion into the plasmid were designed with approximately 350bp homology arms either side of the target mutation. This may also have affected the efficiency of template incorporation, since it is recommended that longer flanking sequences of 500-800bp are used for optimal HDR[340,350,351]. The shorter arms were chosen in this case in order to try to reduce the size of the plasmid, and also to facilitate the cloning process, with PCR fragments of 750bp much easier to clone into the plasmids than fragments of 1,600bp.

For the design of the ssODN template sequences, the homology arms were reduced to 50bp flanking the target sequence, which had previously been demonstrated to enable efficient HDR[469]. Other factors associated with ssODN design, such as using as asymmetric homology arms, could also be used to increase HDR efficiency in the future[470]. These, as well as other possible techniques to improve HDR efficiency are described in more detail in 6.3.3.

5.2 Candidate SNPs Modifying Expression from the β-globin Locus

Two candidate SNPs in KLF1 and ASH1L had previously been identified by Professor Thein's laboratory group. These were identified by independent genetic analyses looking for SNPs causative of an altered pattern of expression from the β-globin locus. This project aimed to establish a CRISPR-Cas9 system for genomic editing in the laboratory, and to use this system to replicate the KLF1 & ASH1L SNPs in the erythroleukaemic K562 cell line. This will allow future functional analysis of each variant *in vitro*, providing insight to any effect that these mutations may be having on globin gene expression in these patients.

5.2.1 KLF1 SNP

This SNP was identified in a currently unpublished genetic study performed by Professor Thein's laboratory. The study performed Sanger sequencing across the full length of the KLF1 gene, including introns, in selected patients from their collection of >800 SCA DNA samples. 50 patients were selected, 25 for having abnormally high HbF, and 25 for having abnormally low HbF, with no known cause. The study hypothesised that as a key regulator of the γ-globin to β-globin switch, novel KLF1 mutations could be causing this strong phenotypic effect. The SNP (rs10407416) was overrepresented in the high HbF group (9/25) compared to the low HbF group (1/25), and is hypothesised to result in the downregulation of KLF1.

This candidate SNP (rs10407416) is an intronic C to G substitution, situated 135bp downstream of exon 1, in an intron of approximately 900bp. This first intron of KLF1 has previously been suggested as a potential downstream regulatory element, and contains highly conserved GATA and SMAD5 binding sites, although these are situated approximately 500bp downstream of the candidate SNP[471]. Reporter assays have also demonstrated that inclusion of intron 1 is required for optimal expression from the KLF1 promoter[471]. The site of the KLF1 candidate SNP

The SNP falls within possible DNA binding sites for transcription factors ZBTB7A and KDM5B (as annotated by data from the ENCODE Consortium[472,473], Figure 5.1), and it is hypothesised that the SNP may disrupt KLF1 expression by interfering with transcription factor binding, resulting in de-repression of γ-globin expression. However, upon further inspection of the ChIP-Seq tracks (Figure 5.1), the signal strength underlying the predicted ZBTB7A binding is relatively weak in K562 cells. KDM5B appears to be strongly associated with the full length of the KLF1 gene in K562 cells, which suggests a strong involvement with the repression of KLF1 in these cells. It is not clear whether binding at the specific site of the SNP would be required to maintain this pattern.

Despite the lack of reliable evidence form the ChIP-Seq data that a specific transcription factor binding at this intronic site significantly affects KLF1 expression, it was decided that the location of a known regulatory element 500bp downstream, as the well as the well-established role that KLF1 plays in globin gene expression, was sufficient to warrant further investigation.

Loss of function of KLF1 has previously been linked to increased HbF levels, and its role in regulation of erythropoiesis and the switch from γ-globin to β-globin expression is well established, and is described more fully in 1.6.1 and 1.3. As such, we hypothesised that this

intronic SNP would reduce KLF1 expression levels, and as a result prevent the efficient switch to β-globin expression, accounting for the increased HbF levels observed in this patient.

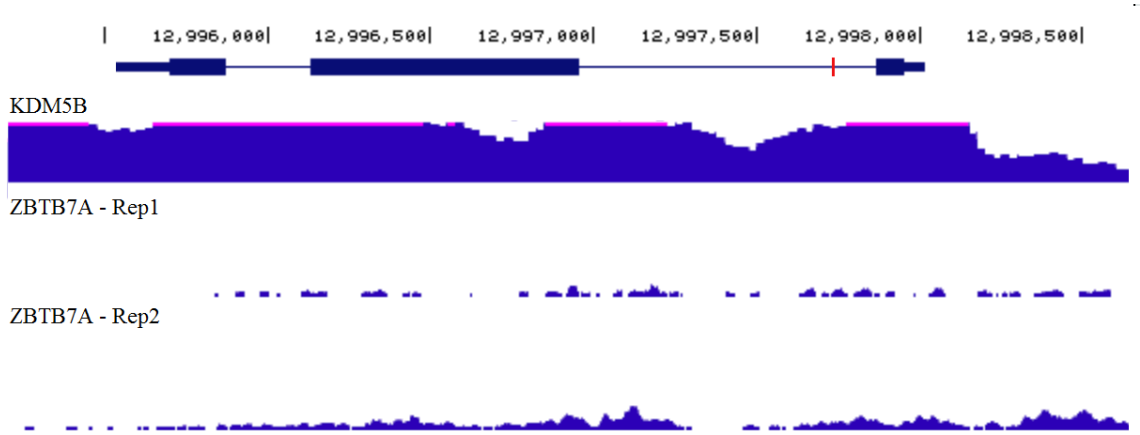


Figure 5.1: Figure showing the full length of the KLF1 gene as viewed in the UCSC Genome Browser (*http://genome.ucsc.edu* - Assembly GRCh37/hg19[380]). Transcription occurs on the negative strand, and the red line indicates the position of the KLF1 SNP (rs10407416) in intron 1. The tracks below show ChIP-Seq signals for KDM5B, as well as two ZBTB7A replicates in K562 cells. It can be seen that there is a strong signal for KDM5B along the length of the gene, but that the signal for ZBTB7A is weak. This data was produced as part of the ENCODE Project[474], and the tracks for KDM5B, and ZBTB7A have UCSC accession numbers wgEncodeEH002085 & wgEncodeEH001620, respectively.

5.2.2 ASH1L SNP

This SNP was identified by a study investigating a large family with heterozygous cases of β-thalassaemia affecting three generations, that did not associate with genetic haplotype at either β-globin or α-globin loci[475,476]. The study had used Whole Genome Sequencing to analyses two affected and two unaffected family members, and identified 15 variants that were then sequenced in the remaining 25 family members[476]. Of these 15 variants, four were present in all seven of the affected family members. Two of these genes were found to be expressed in human erythroid progenitor cells. LRIG2 was expressed at low levels throughout differentiation, whereas increased ASH1L expression occurred shortly before the increase in globin gene expression[476]. The SNP in ASH1L was therefore identified as the most likely candidate to cause the β-thalassaemia phenotype.

ASH1L is a histone methyltransferase that tri-methylates H3K4 at actively transcribed genes, some evidence also suggests that it mono-methylates and di-methylates H3K36[477,478]. Methylated H3K4 and H3K36 act as positive markers for active transcription, and prevent the addition of repressive histone markers such as tri-methylation at H3K27[479]. ASH1L has

previously been shown to occupy promoters at both the α-globin and β-globin like loci, and ASH1L shRNA knock down results in loss of H3K4 tri-methylation and transcription at these promoters in human erythroid progenitors in an *in vitro* culture[476,480]. This suggests a mechanism by which ASH1L is recruited to the promoter and gene body of the β-globin gene, and tri-methylates H3K4, to respectively initiate and protect active transcription[478]. It is therefore plausible that any mutation impairing either the recruitment or catalytic activity of ASH1L would adversely affect β-globin expression, and could cause the observed phenotype.

ASH1L consists of a SET domain, which performs the methyltransferase function, as well as four AT hook motifs, a Bromo-domain and a PHD motif[478,481]. AT hook motifs bind to DNA, while Bromo-domains and PHD motifs are thought to be involved in recognising specific histone modifications (acetylated lysine and methylated lysine respectively), suggesting that ASH1L activity may also be regulated by the chromatin state[482–484]. The candidate SNP identified in ASH1L (rs151028549) is a T to C substitution, that results in Arginine at position 1615 being replaced by Glycine, in a serine rich region. While the SNP is not in close proximity to any of the domains previously associated with recruitment or catalytic function, it was selected for further investigation due to the strength of the evidence for its involvement in β-thalassaemia, as identified by the study.

Since the ASH1L SNP appears to be the strongest candidate for causing β-thalassaemia in this family, we hypothesised that when introduced into K562 cells, this SNP would present a similar effect, and would significantly reduce β-globin expression.

5.2.3 K562 Cells as a model for the KLF1 & ASH1L SNPs

There are several cell systems available for the study of erythropoiesis and globin expression, ranging from primary human erythroid cells extracted from peripheral blood or bone marrow, to established cell lines such as K562. For this project, K562 cells were chosen for the initial functional analyses, since they are well established as a laboratory model, and were thought to be less sensitive to the stressful conditions associated with the CRISPR-Cas9 system. This was a particularly important consideration, since the main aim of this project was to optimise a CRISPR pipeline in the laboratory.

As was described in detail in 1.4.3 and 3.1, it is possible to culture primary erythroid cells from tissue including peripheral blood and bone marrow, and to induce them to differentiate *in vitro*.

While the use of CRISPR-Cas9 to introduce loss of function mutations to these cells has subsequently been demonstrated, this results in a genetically heterogeneous population of cells with a limited lifespan[485]. This would not have been appropriate for this project, and it was anticipated that the length of time required to introduce specific mutations, rather than targeted deletions, would have presented a particular challenge for a short-lived culture. Additionally, given our experience of the sensitivity of these cultures (3.1), it was thought that expansion of clonal cultures from single cells would have been unlikely to be successful.

Induced Pluripotent Stem Cells (iPSCs) expressing the globin genes can be derived from erythroblasts, and would provide a more stable alternative to working with primary cell cultures[486,487]. This would be especially valuable if we were able to generate iPSCs from the erythroblasts of the individual patients themselves, as has recently been demonstrated for a β-thalassaemia patient, where the causative mutation was subsequently corrected *in vitro* using CRSIPR-Cas9[488]. This technique would likely be the most informative on the impact of the SNPs of interest on globin gene expression, and would be worth pursuing in the future. However, due to the technical difficulties associated with generating iPSCs, as well as the high stress associated with the CRISPR-Cas9 system, it was decided that the technique should be optimised in a less sensitive model initially.

Another alternative would be the use of already established immortalised erythroblast lines such as HUDEP-2, and more recently BEL-A, which retain an erythroblast phenotype, and can be induced to differentiate all the way through to terminal erythrocytes[489,490]. Over the last few years, HUDEP-2 has become widely used as an *in vitro* model of globin regulation and erythropoiesis[82,85,491]. HUDEP-2 cells were investigated as a potential model for use in this project, however at the time, our collaborators in Professor Thein's laboratory were experiencing difficulty in reliably culturing these cells, particularly in low cell numbers. It was therefore decided, as with iPSCs, that while this would be very informative, and worth pursuing in the future, the fact that the CRISPR-Cas9 pipeline had not been used in our laboratory before meant that a simpler model should be used first, to optimise the process.

K562 erythroleukaemic cells were chosen as a robust and easy to culture cell line, with active expression form the globin gene loci. K562 is a less accurate model than those discussed above, for example due to the fact that KLF1 expression is very low, and expression of γ-globin is much higher than β-globin, which is more similar to the foetal pattern of globin expression,

that our SNPs of interest are hypothesised to replicate[67,492,493]. Haemoglobin production also low in these cells, however can be increased by induction of differentiation, where the cells acquire a red/pink colour[494]. As a cancer cell line that was generated over 40 years ago, K562 also has an abnormal karyotype, being triploid for most chromosomes (although this karyotyping may not be accurate for our cells, and should be updated), and the relevance of it as a model for erythroid tissues is questionable[390,495].

While KLF1 expression is very low in these cells, it was hypothesised that KLF1 expression may still change in response to disruption of the putative regulatory region. For investigation of the ASH1L SNP, K562 is more appropriate. It had been previously shown that ASH1L binds to the both the α-globin and β-globin promoters in K562 cells, and that this correlates with H3K4me3 at these regions. It was therefore hypothesised that introduction of the ASH1L SNP would disrupt either the recruitment or catalytic activity of ASH1L.

5.3 Transfections & Single Cell Sorting

5.3.1 Nucleofection is the most efficient transfection technique for K562 cells

Three different transfection techniques were tested for efficiency; Lipofectamine, Calcium Phosphate and Nucleofection (2.6.4). K562 cells were sorted into single cell cultures 48 hours after transfection, using GFP as the positive marker for successful transfection. A comparison of the three transfection techniques is shown in Figure 5.2. Efficiency for the Lipofectamine and Calcium Phosphate transfections was very low, ranging from 0.2% to 2.1%. Transfection by Nucleofection was also low, but was comparatively higher with a mean efficiency of 15.2% despite using half the amount of plasmid of the Lipofectamine transfections, and a quarter of that used for the Calcium Phosphate transfections. Different amounts of plasmid were used as a result of the different restrictions on reaction volume for each technique.

All subsequent transfections were therefore performed using the Nucleofection method. Because successfully transfected cells were sorted into single cell cultures for clonal expansion, the low transfection efficiency did not affect the experimental outcome, since a maximum of 288 cultures were grown from each transfection reaction (3 x 96 well plates), and even 0.2% of the initial $1x10^6$ cells provides 2,000 successfully transfected cells.

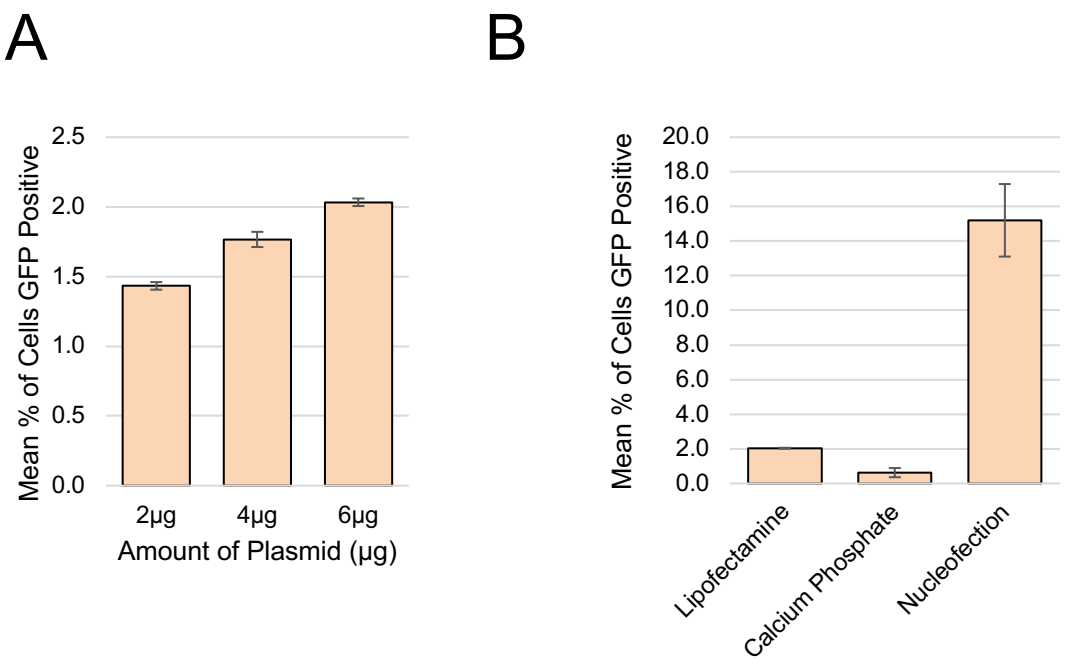


Figure 5.2: Cas9 plasmid transfections in K562 cells. A – Percentage of cells GFP+ 48 hours after Lipofectamine transfection with different amounts of plasmid. Transfection rate increased with increasing concentrations of plasmid, but was very inefficient, reaching only 2% of live cells. B – Percentage of cells GFP+ 48 hours after transfection using the three different techniques. Due to differing restrictions on transfection reaction volume for each technique, different plasmid amounts were used: Lipofectamine - 6µg, Calcium Phosphate - 12µg and Nucleofection - 3µg. Nucleofection was by far the most successful, despite using the least amount of plasmid. Error bars indicate standard error, for each of the Lipofectamine transfections and the Nucleofection n = 3, for Calcium Phosphate n = 4.

5.3.2 Low K562 viability from single cell cultures

After the sorting of GFP+ cells into separate wells by FACS, less than 10% of the clonal cultures survived. This is shown in Figure 5.3, which summarise survival rates from eight different nucleofection experiments, which resulted in the plating of 1,920 single cell sorted cultures, of which only 170 survived and were expanded into clonal cell cultures.

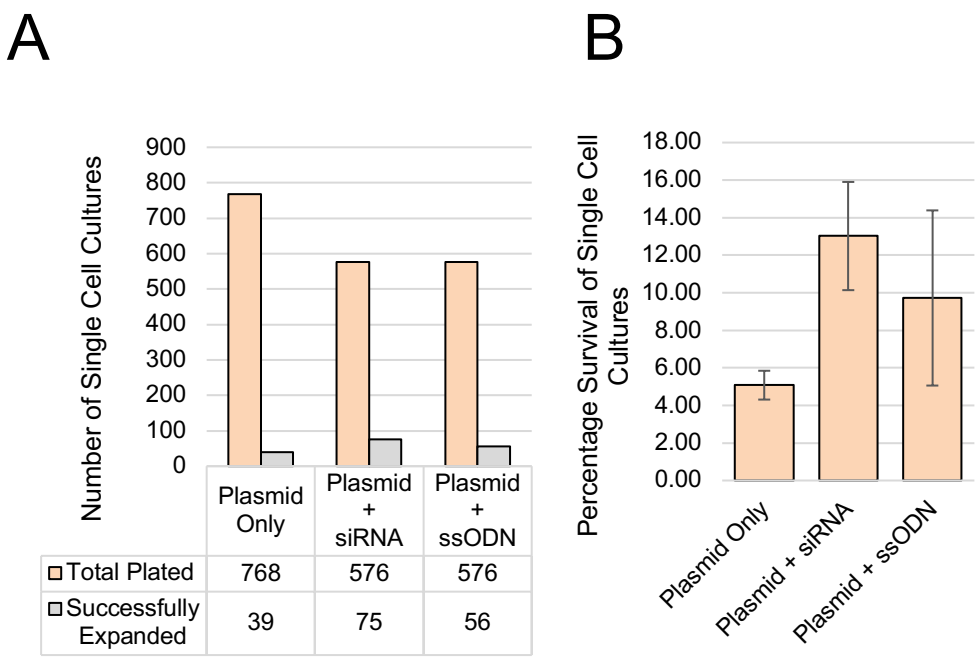


Figure 5.3: Summary of clonal expansions from 12 nucleofection reactions. 4 where only the Cas9-gRNA-Template plasmids were transfected, 6 with the plasmids and siRNA for knockdown of the NHEJ pathway, and 2 with the plasmids and additional ssODN templates. A – Summary of the 1,920 single cell cultures plated, of which only 190 survived. B – Percentage survival for each of the three nucleofection conditions. Survival was low for all experiments, but interestingly was lowest when transfected with the plasmid only. Error bars indicate Standard Error.

This loss in viability could be due to the hydrodynamic stress associated with the FACS process, or as a result of the isolated culture itself, growing in the absence of the growth factors usually secreted into the culture medium when cultured in larger numbers[398,496,497]. The latter could be accounted for by culturing in 'conditioned' medium, whereby the medium from an untransfected K562 culture is removed after 24 hours, filtered to prevent contamination, and then used to grow the single cell cultures.

5.4 Template Incorporation into Genome

5.4.1 CRISPR-Cas9 Cleavage Activity is High, but Template Uptake is Low in K562 Cells

CRISPR-Cas9 experiments were designed to provide the optimal environment for the intended outcome, using PAM site disruptions to prevent repeated cutting once the correct variant has been introduced, and ensuring that the artificial template was more abundant in the cell than the endogenous template, of which there should only be two copies. Despite this, the editing process relies on the stochastic action of Cas9 and the endogenous repair machinery within each cell, and with one million cells per transfection reaction, a variety of genotypes are produced. The broad range of observed genotypes confirms the need to isolate individual cells and culture clonal cell lines that can be analysed individually, rather than producing a heterogeneous mixture of cells with different genotypes.

It is difficult to determine the exact mechanism by which any given genotype has been introduced into a cell using CRISPR. For example, a cell line homozygous for the desired variant may have undergone successful cleavage and template introduction independently on both alleles, or only on one allele which then acted as a template for the second. Similarly, for cell lines homozygous for the wild type allele, it is possible that no cleavage took place at all, or that one allele was cleaved, but that the other allele was used as the HDR template.

In order to assess the efficiency of the CRISPR-Cas9 experiments based on the genotypes produced, some assumptions were made. Firstly, that any cell line with the wild type genotype experienced no Cas9 cleavage. Secondly, any cell lines homozygous for a variant underwent HDR. And thirdly, any allele other than the wild type or the template was produced as a result of NHEJ.

The results of the genetic screening for successful CRISPR cell lines after transfection with the CRISPR-Cas9 plasmids and expansion from single cell cultures are shown in Figure 5.4. These results illustrate that Cas9 is being directed to the correct genomic loci and cleaving efficiently, with genetic variants introduced in 79.5% of the cell lines screened. However, introduction of the SNPs of interest was much less successful, with only 25.8% of these variants including an allele matching the desired mutation, and none of these were homozygous.

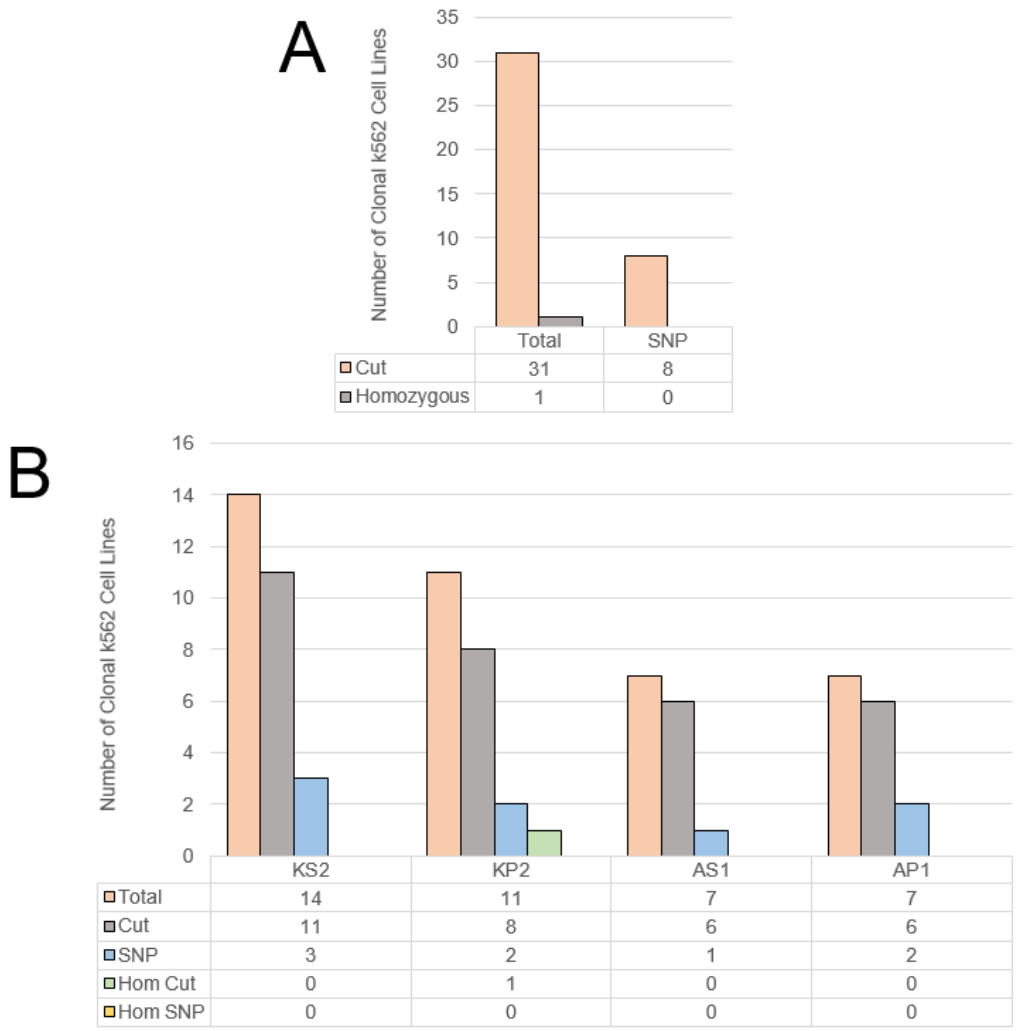


Figure 5.4: Summary of genetic analyses of K562 cell lines after transfection with CRISPR-Cas9 Template containing plasmids only, after subsequent FACS and clonal expansion. A – Summarises the results for all plasmids. B – Shows the results for each plasmid individually. Plasmids used were for KLF1 gRNA 2, SNP and PAM only control (KS2 & KP2 respectively), and ASH1L gRNA 1, SNP and PAM only control (AS1 & AP1 respectively). Total refers to the number of cell lines that survived the single cell sorting stage. Cut refers to cell lines where any genetic changes have occurred, SNP refers to cell lines where the template mutations have been introduced on any allele, Hom Cut or SNP refers to cell lines defined as Cut or SNP that are homozygous. The results show that the gRNA-Cas9 plasmids cut with high efficiency, but introduction of the template is much less successful. Only one cell line was homozygous for a genetic variant, and none were homozygous for the SNPs of interest.

Only one cell line was homozygous for a genetic modification introduced by CRISPR, equating to 2.6% of all the cell lines screened. This variant was a 10bp deletion at the gRNA cleavage site, and the majority of variants observed in all cell lines were also short insertions or deletions. These variants are presumably the result of the NHEJ pathway, and are likely so prevalent due to the fact that the products are very stable. Using the same rationale as for the PAM site disruption mutations, these insertions or deletions either disrupt the PAM site or remove it completely. Some deletions also result in the removal of the gRNA target sequence, and are very efficient at preventing repeated cleavage by Cas9.

The results suggest that NHEJ is occurring much more frequently than HDR in K562 cells, this is demonstrated not only by the frequency of insertions and deletions, but by how few of the cell lines had homozygous mutations, suggesting that it is rare for the already modified allele to be used as a template.

This highlights one of the main limitations of using the CRISPR machinery for targeted genetic editing. While the introduction of DSB by Cas9 is efficient, the endogenous repair machinery is relied on to introduce the target mutations, and the efficiency of this varies between cell types depending on the activity of the NHEJ and HDR pathways.

5.4.2 siRNA Mediated Knockdown of NHEJ pathway

Template sequence uptake in K562 cell lines after cleavage by Cas9 was very low. This was thought to be due to high activity of the NHEJ pathway, with fewer double strand breaks being repaired by the HDR pathway that is required to incorporate the repair template into the genome. It was hypothesised that by inhibiting the NHEJ pathway, increased repair through the HDR pathway would be observed. This was tested by siRNA mediated knockdown of components of the NHEJ machinery, with siRNA transfected simultaneously with the CRISPR-Cas9 and template containing plasmid.

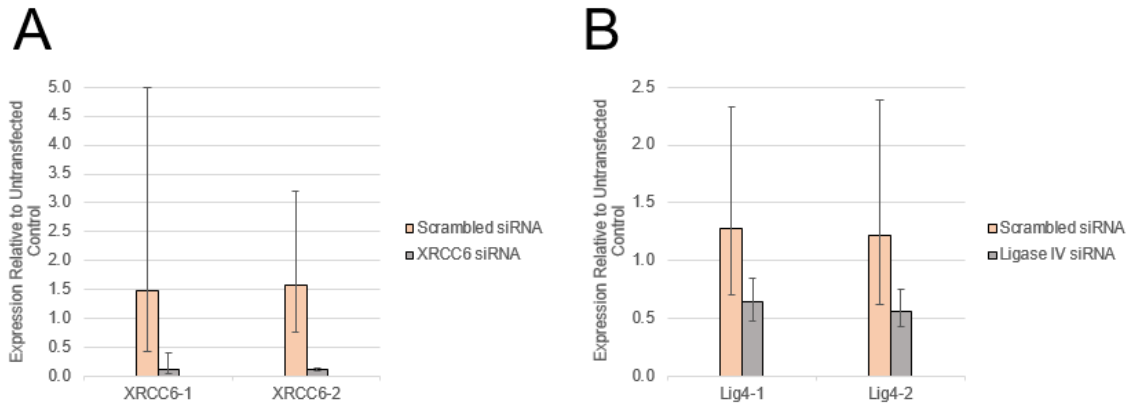


Figure 5.5: rtPCR analysis of NHEJ knockdown by siRNA in K562 cells, normalised firstly to β-actin expression, and then to the untransfected control. rtPCR analysis was performed on RNA extracted 48 hours after transfection with either scrambled siRNA or targeted siRNA. A – Knockdown using siRNA for XRCC6. B – Knockdown using siRNA for Ligase IV. Results show reduced expression for both XRCC6 and Ligase IV, 11.6% and 60.2% of untransfected K562 expression respectively. Expression appears to have increased in the scrambled controls, although large variation was observed. Two sets of PCR primer pairs were used for each gene targeted, XRCC6-1 & 2 and Lig4-1 & 2, and results are consistent between each pair. Error bars indicate 95% confidence intervals, calculated from three biological replicates, each with two technical replicates. Knockdown of XRCC6 was statistically significant compared to scrambled, whereas Ligase IV was not, likely due to the variation observed between the samples transfected with scrambled siRNA.

In order to inhibit the NHEJ pathway, either the first or the last component of the pathway were knocked down, (XRCC6 or Ligase IV respectively). XRCC6 (X-Ray Repair Cross Complementing 6, aka Ku70) forms a dimer with Ku80, and this complex recognises and binds the dsDNA ends and acts as a scaffold to recruit the other factors of the NHEJ pathway[498–502]. Ligase IV is recruited to the NHEJ complex at the double strand break through interaction with XRCC4, and performs the ligation step, re-joining the two ends of the double strand break[502,503].

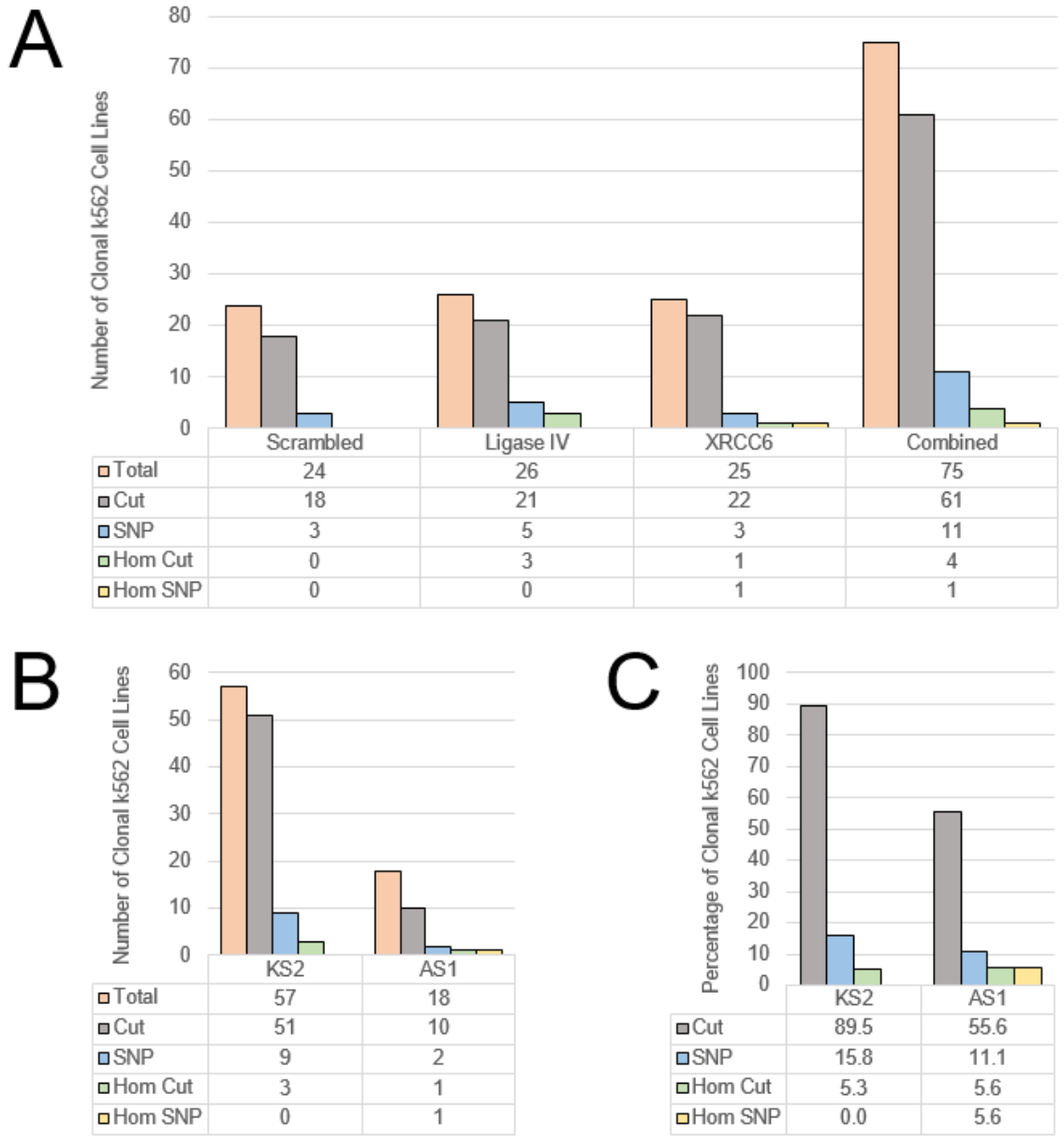
The results in Figure 5.5 show that knockdown of XRCC6 & Ligase IV by siRNA was successful in K562 cells. Cells transfected with siRNA targeting XRCC6 demonstrated a roughly 90% reduction in XRCC6 expression, while cells transfected with siRNA targeting Ligase IV had a less efficient knockdown of roughly 40%, and did not reach statistical significance. The scrambled siRNA controls showed were highly variable for both genes assayed.

Having tested and confirmed the successful knockdown of XRCC6 and Ligase IV by siRNA, the CRISPR-Cas9 plasmids containing the SNP templates were co-transfected with XRCC6, Ligase IV or scrambled siRNA. The results of the genetic screening of clonal cell lines arising from these transfections are summarised in Figure 5.6.

It is not clear why the error associated with the scrambled siRNA transfected controls in Figure 5.5 is so high. Given the small degree of variation observed in the targeted siRNA cultures, and the fact that two pairs of target primers were used, it seems unlikely that this is an artefact of the rt-PCR itself. It has been shown that non-targeting siRNA can trigger a stress response within the cell, which may account for some of the variation that is observed[504]. If these siRNA experiments are used to reduces NHEJ activity in the future, western blotting analysis should be run in parallel with the rt-PCR to investigate this variation and confirm that the changes in expression are observed at the protein level.

The results in Figure 5.6 show that Cas9 cleavage activity is not reduced in cell lines co-transfected with siRNA targeting either Ligase IV or XRCC6, with 75.0% containing a genetic variant when transfected with scrambled siRNA, 80.1% for Ligase IV and 88% for XRCC6, with an average efficiency of 81.3%. Compared to the 79.5% observed for the data shown in Figure 5.4 for cells transfected with the CRISPR-Cas9 plasmid only, this demonstrates that siRNA transfection does not alter Cas9 activity, despite the fact that limitations to the transfection volume mean that half as much plasmid is used during the co-transfection experiments. This shows that while the amount of plasmid used for transfection is important for transfection efficiency, it has a limited effect on Cas9 activity within the cell.

This is important to take into consideration when designing experiments to generate clonal cell lines, where only a few hundred individual cells are positively selected for downstream culturing, making transfection efficiency largely irrelevant.



	Scrambled	Ligase IV	XRCC6	Combined
Total	24	26	25	75
Cut	18	21	22	61
SNP	3	5	3	11
Hom Cut	0	3	1	4
Hom SNP	0	0	1	1

	KS2	AS1
Total	57	18
Cut	51	10
SNP	9	2
Hom Cut	3	1
Hom SNP	0	1

	KS2	AS1
Cut	89.5	55.6
SNP	15.8	11.1
Hom Cut	5.3	5.6
Hom SNP	0.0	5.6

Figure 5.6: Summary of genetic analyses of K562 cell lines after transfection with CRISPR-Cas9 Template containing plasmids and siRNA, after subsequent FACS and clonal expansion. Total refers to the number of cell lines that survived the single cell sorting stage. Cut refers to cell lines where any genetic changes have occurred, SNP refers to cell lines where the template mutations have been introduced on any allele, Hom Cut or SNP refers to cell lines defined as Cut or SNP that are homozygous. A – Summary of the cell lines transfected with each siRNA set: Scrambled, Ligase IV or XRCC6, as well as the cumulative counts for all three. B & C – Summary of the cell lines transfected with either KS2 or AS1 plasmids, B shows total counts, C shows percentage of total. Results show that co-transfection with siRNA for one of the target genes does not appear to affect Cas9 cutting activity, which is consistent between the three groups. No homozygous variants were observed after transfection with scrambled siRNA, whereas three were observed with siRNA targeting Ligase IV, and one for XRCC6. Overall survival of cell lines past the single cell FACS stage is much higher for the KS2 plasmid than for AS1. One of the AS1 cell lines (KAX9) was found to be homozygous for the desired SNP.

Interestingly, it appears that cells transfected with the AS1 plasmid have a much lower survival rate than for the KS2 plasmid, with 6.3% and 19.8% of single cell sorted cultures surviving

respectively. This was also observed in the plasmid only transfections shown in Figure 5.4, where the two K2 gRNA containing plasmids, KS2 & KP2 had survival rates of 7.3% and 5.7% respectively, compared to the A2 gRNA containing plasmids, which both had 3.6% survival. Similarly, a higher percentage of KS2 cell lines showed Cas9 activity (89.5% vs 55.6% for AS2), however this was not observed in the plasmid only transfections.

While it was shown that siRNA did not impair Cas9 activity, it remains unclear as to whether or not knocking down the Ligase IV or XRCC6 influenced the number of homozygous variants that were produced, since the success rate was still extremely low. No homozygous variants were generated using the scrambled siRNA, compared to 3 for the Ligase IV siRNA and 1 for the XRCC6 siRNA, making the HDR rates for each 0.0%, 11.5% and 4.0% respectively, compared to 2.6% for the plasmid only transfections from Figure 5.4. A Poisson Test was performed on these success rates, using 0.026 (from the plasmid only transfections) as the expected mean. The results of this test are shown in Table 5.1, and indicate that siRNA mediated knockdown of Ligase IV resulted in a significant increase in the success rate of generating homozygous genetic variants in K562 cells. However, due to the extremely low numbers involved in calculation of the mean success rates, these tests may not be reliable

Null Hypothesis: $\lambda = 0.026$	Scrambled	Ligase IV	XRCC6
Number of Cell Lines (n)	24	26	25
Expected Successes ($X = n \times \lambda$)	0.624	0.676	0.65
Observed Successes (O)	0	3	1
P(O = X)	**0.5358**	**0.0262**	**0.3393**

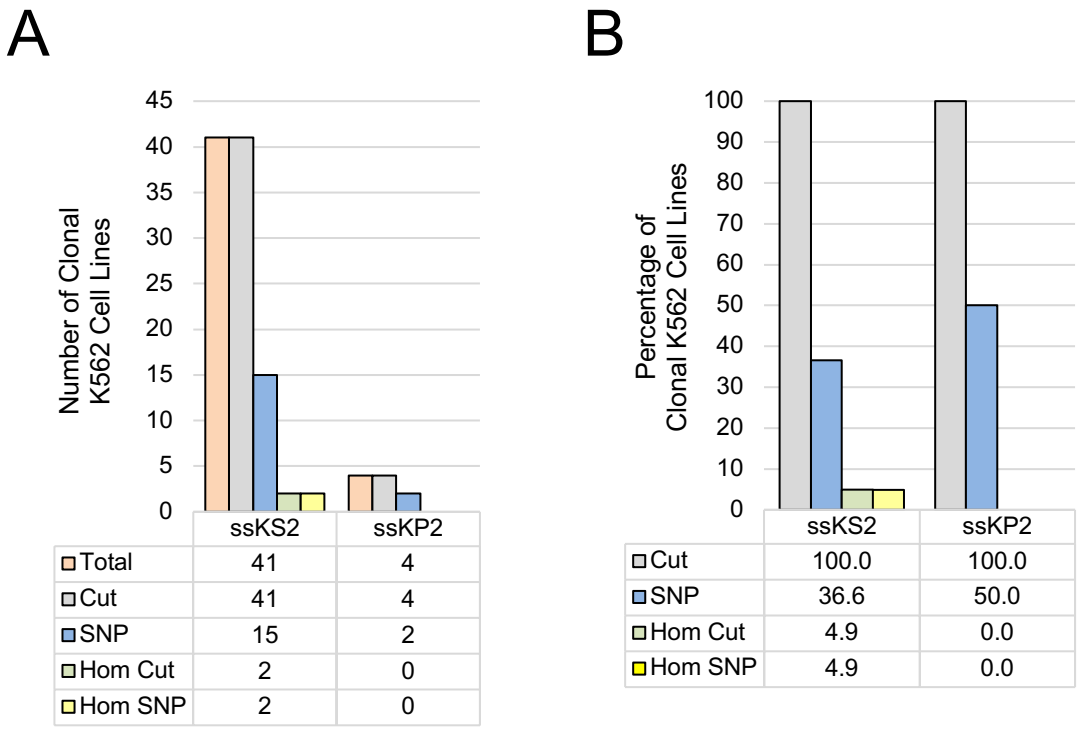
Table 5.1: Poisson test for significance for the increase in success rate when generating homozygous genetic variants using siRNA for Ligase IV or XRCC6. Probability was calculated using the Poisson Distribution Calculator made available online at ncalculators.com[505]. K562 cells transfected with siRNA targeting Ligase IV were the only group able to reject the null hypothesis at the significance threshold of $p<0.05$.

Knocking down Ligase IV yielded three homozygous variants for the KS2 plasmid, however these were all deletions, and did not contain the SNP of interest. Knocking down XRCC6 yielded one homozygous variant for the AS1 plasmid, which was found to be homozygous for the variant of interest, and therefore can be confirmed to be a successful introduction of the ASH1L SNP into K562 cells using the CRISPR-Cas9 system.

5.4.3 ssODN to Increase Template Copy Number in the Cell

While the results of knocking down Ligase IV may have shown a significant increase in the number of homologous genetic variants generated, the success rate was still very low, and the rate of successful template incorporation into the genome was even lower. It was thought that this was possibly due to the low copy number of the template in the cell, as a result of the low transfection efficiency.

Reducing the amount of plasmid used for transfection was shown to not affect Cas9 activity, but reducing the amount of template may affect HDR efficiency. Having the template incorporated into the CRISPR-Cas9 plasmid allows selection for successful transfection using GFP expression, however it limits the intracellular levels of template to a ratio of 1:1 with the larger plasmid.



	ssKS2	ssKP2
Total	41	4
Cut	41	4
SNP	15	2
Hom Cut	2	0
Hom SNP	2	0

	ssKS2	ssKP2
Cut	100.0	100.0
SNP	36.6	50.0
Hom Cut	4.9	0.0
Hom SNP	4.9	0.0

Figure 5.7: Summary of genetic analyses of K562 cell lines after transfection with CRISPR-Cas9 Template containing plasmids and ssODN templates for KS2 or KP2 (ssKS2 or ssKP2), after subsequent FACS and clonal expansion. Total refers to the number of cell lines that survived the single cell sorting stage. Cut refers to cell lines where any sequence changes have occurred, SNP refers to cell lines where the template mutations have been introduced on any allele, Hom Cut or Hom SNP refers to cell lines defined as Cut or SNP that are homozygous. A – Total cell line counts. B – Percentage of total. Cleavage was observed in all cell lines, and SNP uptake was high. Number of homozygous cell lines remained low, however two ssKS2 cell lines were homozygous for the SNP of interest (ssKS2-10 & ssKS2-29).

It was hypothesised that introducing more copies of the template may increase HDR efficiency. Short 110bp ssODN templates were used to test this, and were transfected alongside the CRISPR-Cas9 plasmids which also contained the template sequences. These short ssODNs

have previously been used as a template for HDR alongside Zinc Finger Nuclease (ZFN) genomic editing techniques[339,347], as well as with the CRISPR-Cas9 system[384].

Since a successful K562 cell line containing the ASH1L SNP had already been generated by this stage, only transfections targeting the generation of the KLF1 SNP were carried out. A summary of the cell lines generated by these transfections is shown in Figure 5.7, and shows that very low viability was observed for the ssODN + KP2 plasmid (ssKP2) transfections compared to ssODN + KS2 plasmid (ssKS2), with only 4 of 288 single cell cultures surviving compared to 41 of 288 for ssKS2. Interestingly, there was a 100% cleavage rate for the cell lines co-transfected with ssODN, and template uptake was observed in 36.6% and 50.0% of clones for ssKS2 and ssKP2 respectively.

Homozygous variant introduction was still low, with only two homozygous clones. However these two clones were homozygous for the SNP of interest, and confirmed successful introduction of the KLF1 SNP into K562 cells using the CRISPR-Cas9 system.

5.5 ASH1L mutant K562 cell line KAX9

5.5.1 Genotype of the K562 ASH1L mutant KAX9

The cell lines that survived the single cell sorting process were screened by Sanger sequencing over the ASH1L SNP site. One successful homozygous cell line was generated for the ASH1L SNP (KAX9, Figure 5.8), which was produced by co-transfection of the AS1 plasmid with the siRNA for XRCC6.

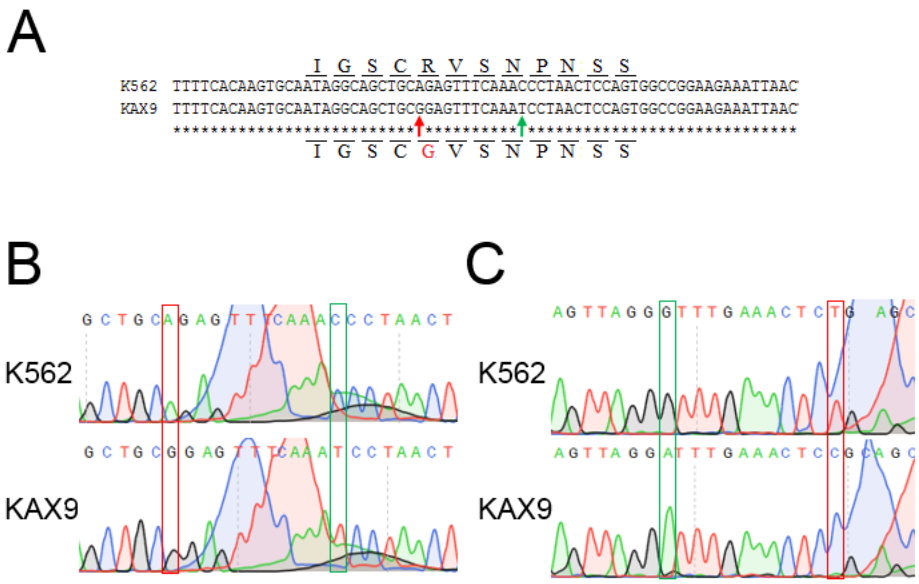


Figure 5.8: Sequence of the ASH1L SNP site of the K562 cell line that was homozygous for PAM disruption mutation and the SNP. K562 shows the wild type untransfected sequence. The green box/arrow shows the site of the C to T PAM disruption mutation. The red box/arrow shows the site of the A to G SNP of interest. A – MUSCLE alignment of the two sequences, with coding sequence displayed (antisense). The two SNPs can clearly be seen in the KAX9 sequence, and it can be seen that the SNP results in an arginine to Glycine substitution, while the PAM disruption does not affect the coding sequence. One other polymorphism was identified, but by investigating the sequence traces was confirmed to be an artefact of the base calling algorithm. B & C – Forward and Reverse sequence traces respectively. Due to the presence of large Sanger sequencing artefacts, that persisted despite repeated sequencing, both forward and reverse sequence traces are shown, to confirm that the both the PAM disruption and SNP are present.

Sequencing artefacts like those in Figure 5.8 are often seen near the start of sequence traces, where large amounts of unbound fluorescent ddNTPs can cause large distortions, however this is not believed to be the case here. The site shown is approximately 80bp downstream from each sequencing primer, and the sequences both upstream and downstream of this region are free from these artefacts for both forward and reverse traces. The fact that these artefacts were observed when sequencing was repeated suggests that this is specific to the sequence and region, and the fact that it was observed in all cell lines assayed, including the K562 wild type controls, confirms that it is not a result of genomic disruption caused by Cas9 cleavage. While it is possible that it could be possible to generate cleaner chromatograms by testing different primer pairs, it is not necessary, and the clear presence of both the SNP and PAM disruption

mutation when sequenced in both directions (Figure 5.8) is sufficient to confirm correct incorporation of the template sequence.

5.5.2 rtPCR Analysis of the K562 ASH1L mutant KAX9

Expression of α-globin, β-globin and γ-globin were found to be significantly increased in the KAX9 cell line containing the ASH1L SNP compared to the wild type K562 cells (Figure 5.9). This was unexpected, given that the SNP was believed to be causative of β-thalassaemia, and it was predicted that K562 cells harbouring the mutation would show reduced expression of β-globin. KLF1 expression remained unchanged in KAX9 compared to K562, suggesting that this global upregulation of globin gene expression was independent of the KLF1 regulatory pathway. A previous study investigating shRNA mediated knockdown of ASH1L in human erythroid progenitors demonstrated a decrease in globin gene expression, having the opposite effect of what was observed in the K562 experiments[476]. This decrease in globin expression correlated with reduced occupancy of ASH1L at the β-globin and α-globin promoters, and a reduction in H3K4 tri-methylation at these regions[476].

It is possible that the change in globin expression occurred as the result of the stressful CRISPR-Cas9 modification process, which involves high pressure FACS, single cell culturing and the introduction of double strand breaks into the genome. The stress involved in these processes is demonstrated by the low survival rates observed in 5.3.2, and it has been shown that K562 cells can be induced to differentiate under conditions of stress[506,507]. During stress induced differentiation, upregulation of all the globin genes is known to occur. However, this increase in globin expression is also accompanied by an increase in KLF1, which was not observed in the KAX9 cells (Figure 5.9)[506]. Similarly, there was no observed change in the colour of the KAX9 cell pellet, which for K562 cells turns a pink/red colour when differentiation is induced, as a result of increased haemoglobin production[494]. Due to the absence of these established markers of K562 differentiation, it is possible that the observed changes in globin gene expression are occurring directly as a result of the ASH1L SNP, since in K562 cells ASH1L is known to bind and tri-methylate H3K4 at promoters at both the α-globin and β-globin loci[476].

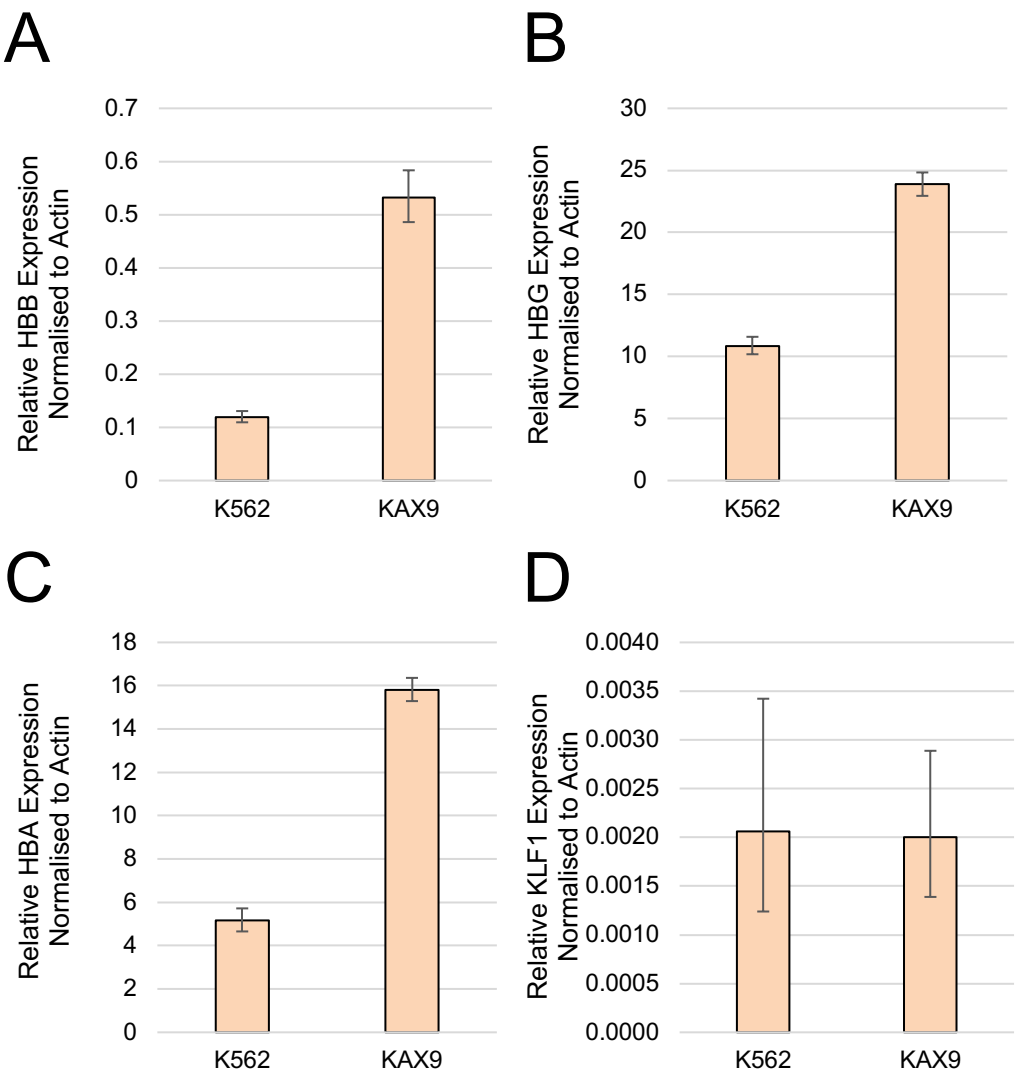


Figure 5.9: rtPCR analyses of wt K562 and KAX9 cell lines. Graphs show relative expression of genes normalised to actin β, for A – β-globin (HBB), B – γ-globin (HBG), C – α-globin (HBA) and D – KLF1. Error bars indicate 95% confidence intervals, calculated from three technical replicates for each of the two cell lines. Expression of the globin genes is significantly increased in KAX9 compared to K562, and KLF1 expression is unchanged.

If the cells are being induced to undergo differentiation, either as a direct result of the SNP, or due to stress associated with the genomic editing process (discussed in 5.3), then this could account for some of the observed changes, since β-actin levels have been shown to decrease during erythroblast development[508]. A relative increase in globin gene expression would be observed if the levels of β-actin RNA were reduced in the KAX9 sample, this could either be a direct effect of the ASH1L SNP, a decrease triggered by differentiation in response to stress, as discussed above, or even due to degradation of the RNA sample. RNA degradation impairs PCR efficiency, and can drastically affect PCR based quantification assays[509]. Quality of RNA samples should be assessed in future to rule this out as a potential cause of the observed

affects, this could be assayed using a variety of techniques, including agarose gel electrophoresis, NanoDrop, or more sophisticated analysis by Agilent's Bioanalyzer[509].

With general upregulation of the globin genes, it is difficult to directly detect any changes in the levels of β-globin and γ-globin that may be occurring as a result of the ASH1L SNP. To account for this, relative expression levels were normalised to α-globin instead of actin (Figure 5.10). While α-globin expression was observed to change between the K562 wild type and KAX9 cells (Figure 5.9), and it is generally not appropriate to normalise to a gene with variable expression, we believe that it is informative in this case. Since expression was increased in each of the globin genes, analysis of the changes in expression relative to each other provides insight into changes in patterns of expression from the loci separately from the general increase in expression that was observed.

It is worth noting that the other genes at the α-globin and β-globin loci were not assayed in these experiments, and may provide useful insight into the pattern of expression changes occurring as a result of the ASH1L SNP. If general disruption of transcriptional regulation is occurring at these loci, as described above, then it would be expected that the transcriptionally repressed ζ-globin at the α-globin locus, as well as ε-globin and perhaps even δ-globin at the β-globin locus would also have increased expression relative to β-globin.

When normalised to α-globin levels, it appears that relative expression of β-globin was increased in KAX9 compared to K562, whilst γ-globin levels were decreased. This is further demonstrated by the ratio of γ-globin:β-globin, which is reduced by approximately 50% in KAX9, indicating a shift towards β-globin expression from γ-globin in these cells.

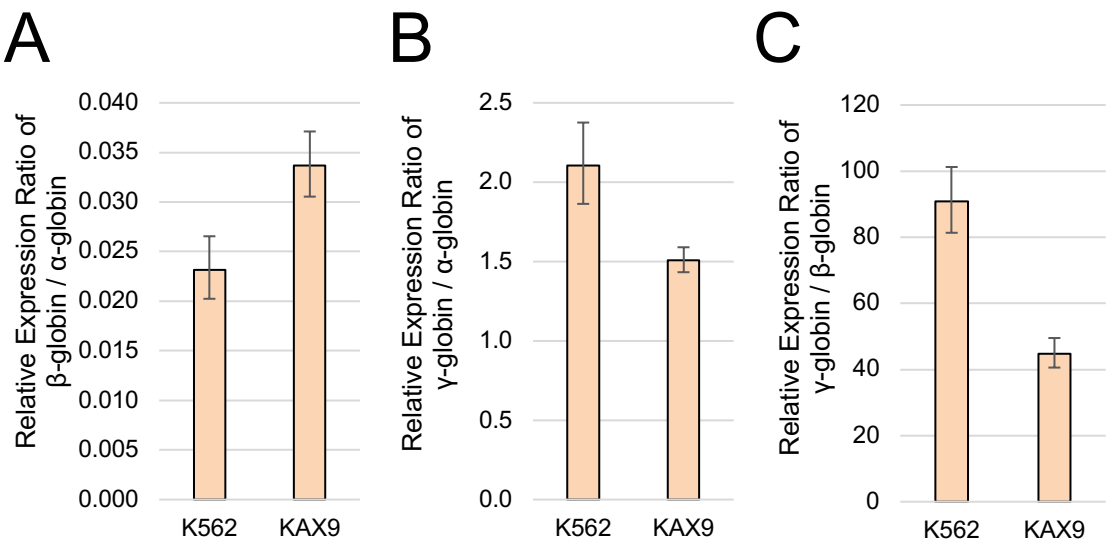


Figure 5.10: rtPCR analyses of wt K562 and KAX9 cell lines, normalised to either α-globin or β-globin expression. A – β-globin normalised to α-globin, B – γ-globin normalised to α-globin, C – γ-globin normalised to β-globin. Error bars indicate 95% confidence intervals, calculated from three technical replicates for each of the two cell lines. Results indicate that relative to α-globin, β-globin increased and γ-globin decreased in KAX9 compared to wt K562. The ratio of γ-globin to β-globin transcripts also decreased in KAX9 cells.

One explanation for the increased expression of β-globin in proportion to γ-globin, rather than the reverse, which is hypothesised to occur *in vivo*, is that the wild type K562 cell line strongly expresses γ-globin expression, while adult human erythroid cells strongly express β-globin. Under these different regulatory states, transcription at the β-globin locus could be affected differently by the ASH1L SNP. This would suggest that rather than causing a specific reduction in β-globin expression, a more general disruption of transcriptional regulation is observed. This is described in more detail in the discussion in 6.3.4.

5.6 KLF1 mutant K562 cell lines

5.6.1 KLF1 CRISPR modified genotypes

5.6.1.1 KLF1 SNP introduction

The cell lines that survived the single cell sorting process were screened by Sanger sequencing over the KLF1 SNP site. Two cell lines were found to be homozygous for the KLF1 SNP (ssKS2-10 & ssKS2-29, Figure 5.11).

Because of the nature of the SNP, and the fact that it is thought to interfere with transcriptional regulation through disruption of a DNA binding site, other cell lines with potentially disruptive genotypes were also selected to test this hypothesis. These include ssKS2-47, which was heterozygous for the SNP and the wild type allele, which is also shown in Figure 5.11. If transcriptional regulation is impaired on one allele, it would be expected that ssKS2-47 would present a phenotype similar to that observed in the homozygotes, but with a weaker effect size.

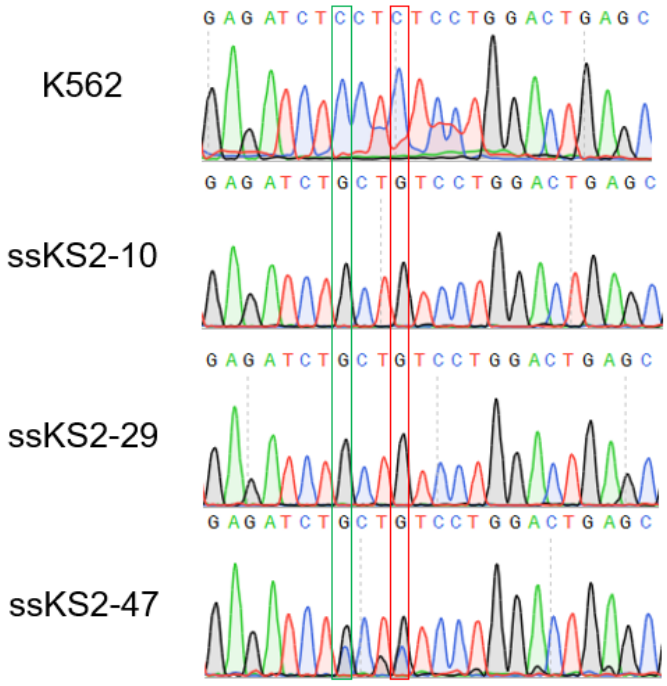


Figure 5.11: Sanger sequencing trances of the KLF1 SNP site of K562 cell lines that incorporated the template sequence on at least one allele, and had no indel mutations on either allele. K562 shows the wild type untransfected sequence. ssKS2-10 and ssKS2-29 were homozygous for both the C to G PAM disruption (green box) and the C to G SNP of interest (red box). ssKS2-47 was heterozygous for both the PAM disruption and the KLF1 SNP.

Three other heterozygous cell lines were also investigated: ssKS2-3, ssKS2-4 & ssKS2-45, all of which contained indel mutations (Figure 5.12). ssKS2-3 was heterozygous for the PAM disruption and SNP of interest, and the other allele contained a CT dinucleotide insertion 2bp downstream from the SNP site. This dinucleotide insertion is situated within the gRNA target

sequence, and likely disrupted gRNA binding, preventing repeat cleavage and incorporation of the target sequence. ssKS2-4 had the same dinucleotide insertion as was observed in ssKS2-3, but the other allele was wild type. ssKS2-4 was chosen due to the similarity in genotype with ssKS2-3, with the only difference being heterozygosity for the SNP of interest. It was thought that ssKS2-4 could be an informative control.

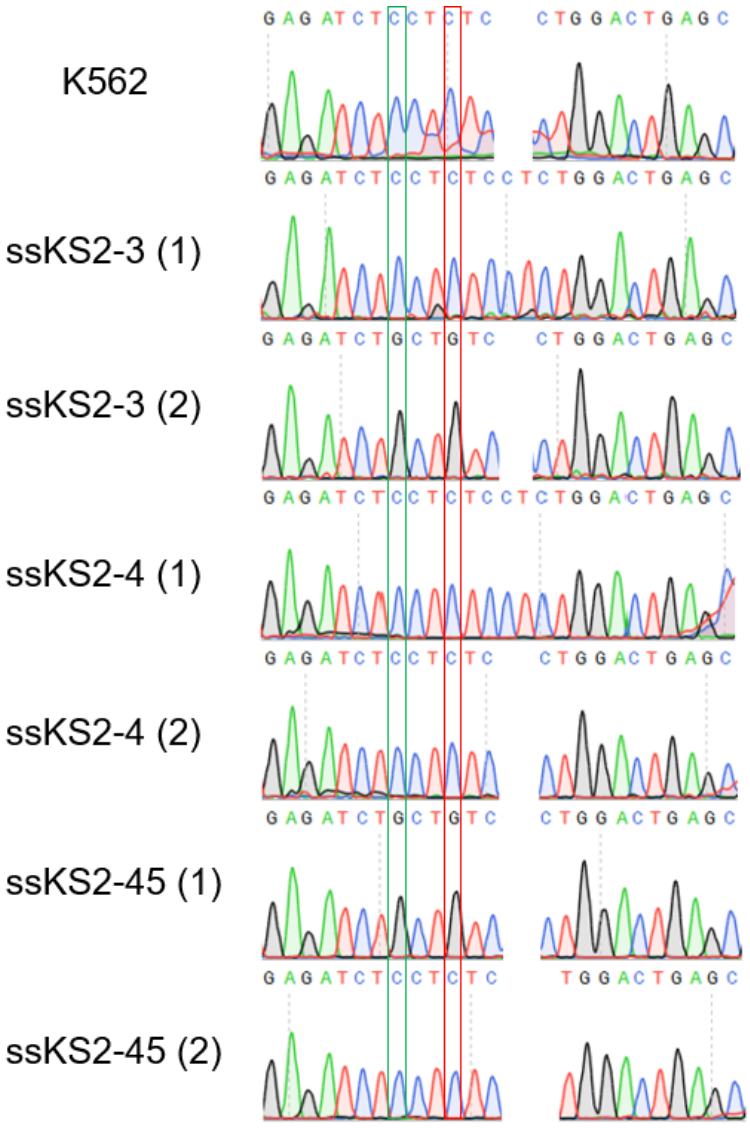


Figure 5.12: Sanger sequencing trances of the KLF1 SNP site of three K562 cell lines that contained heterozygous indel mutations. K562 shows the wild type untransfected sequence. The green box shows the site of the C to G PAM disruption mutation. The red box shows the site of the C to G SNP of interest. Indel mutations prevent clear reading of the sequence from Sanger sequencing traces, since the two alleles are out of frame of each other. Therefore to fully characterise the genotypes of these cell lines, PCR amplicons were cloned and sequenced individually. (1) and (2) refer to two separate alleles for each cell line. ssKS2-3 and ssKS2-4 are both heterozygous for a dinucleotide insertion, and ssKS2-3 has the PAM disruption and SNP of interest on the other allele. ssKS2-45 is heterozygous for a 1bp deletion, and has the PAM disruption and SNP of interest on the other allele.

ssKS2-45 was heterozygous for the PAM disruption and the SNP of interest, and the other allele contained a 1bp deletion of a cytosine 2bp downstream from the SNP site, and likely

disrupted gRNA binding similar to the mechanism suggested for the dinucleotide insertion in ssKS2-3.

The three homozygous cell lines that were generated by co-transfection of the KS2 plasmid and Ligase IV siRNA were also investigated (KKL8, KKL11 & KKL17, Figure 5.13). Although none of these incorporated the template DNA, they each introduced indel mutations around the SNP site. KKL8 was homozygous for a 5bp deletion that covered the SNP site, starting immediately downstream of the site of the PAM disruption mutation. KKL11 was homozygous for the same CT dinucleotide insertion that was heterozygous in ssKS2-3 and ssKS2-4. KKL17 was homozygous for a 41bp deletion that extended 18bp upstream of the site of the PAM disruption mutation, and 19bp downstream of the SNP site. If the region is required for transcriptional regulation, the mutation in KKL17 would be expected to disrupt it.

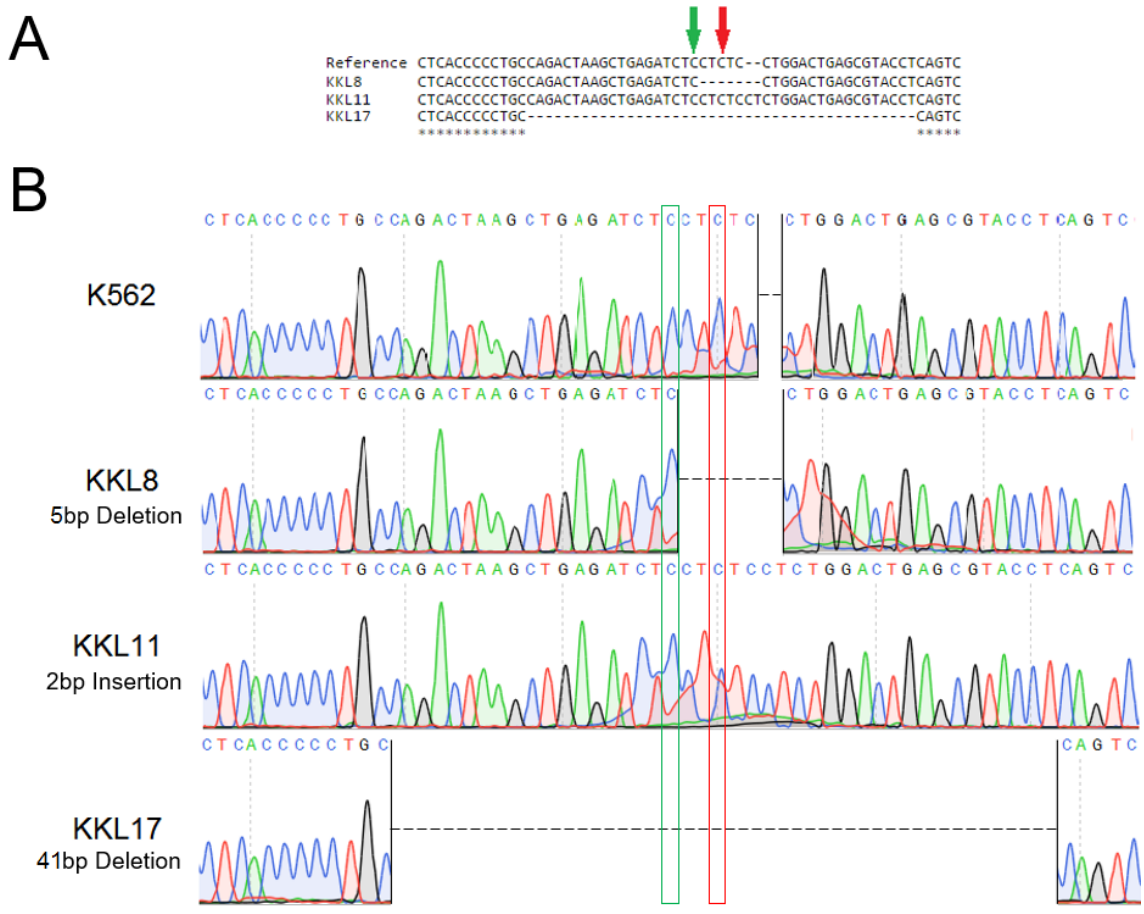


Figure 5.13: Sequences of the KLF1 SNP site of three K562 cell lines that contained homozygous indel mutations. K562 shows the wild type untransfected sequence. The green box/arrow shows the site of the C to G PAM disruption mutation. The red box/arrow shows the site of the C to G SNP of interest. A – MUSCLE alignment of the 4 sequences. B – Sanger sequencing traces. KKL8 was homozygous for a 5bp deletion removing the SNP site. KKL11 was homozygous for a 2bp insertion 2bp downstream of the SNP site. KKL17 was homozygous for a 41bp deletion covering the PAM site and the SNP.

5.6.1.2 KLF1 PAM Disruption Only

While successful K562 cell lines were generated containing the KLF1 SNP and the corresponding PAM site disruption, none of the KP2 plasmid transfections produced cell lines homozygous for the PAM disruption only. Two heterozygous cell lines were generated (ssKP2-3 & ssKP2-4, Figure 5.14), both of which had indel mutations on the other allele. The ssKP2-3 wild type allele also contained a cytosine insertion 1bp downstream from the SNP site, and ssKP2-4 was heterozygous for a 12bp deletion that covered the SNP site and the PAM site. These indel mutations both disrupt the gRNA binding sequence. While these may not be as relevant as homozygous controls, they could be informative when compared to the cell lines heterozygous for the KLF1 SNP.

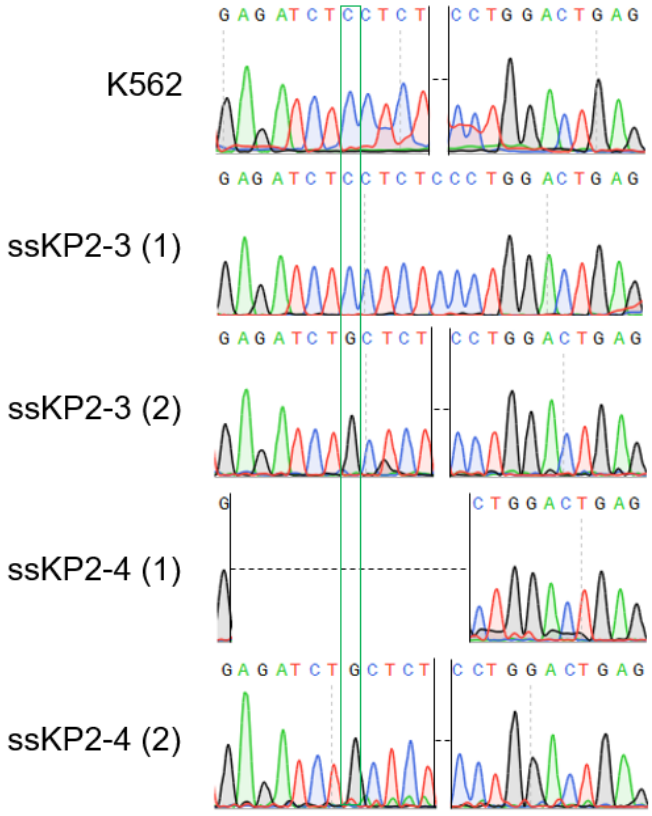


Figure 5.14: Sanger sequencing trances of the KLF1 SNP site of two candidate PAM disruption only controls. K562 shows the wild type untransfected sequence. The green box shows the site of the C to G PAM disruption mutation. Indel mutations prevent clear reading of the sequence from Sanger sequencing traces, since the two alleles are out of frame of each other. Therefore to fully characterise the genotypes of these cell lines, PCR amplicons were cloned and sequenced individually. (1) and (2) refer to two separate alleles for each cell line. ssKP2-3 is heterozygous for the PAM disruption and a 1bp insertion. ssKP2-4 is heterozygous for the PAM disruption and a 12bp deletion.

5.6.2 rtPCR Analysis of KLF1 mutant K562 cell lines

5.6.2.1 KLF1 Expression in K562 KLF1 mutants

The KLF1 intronic SNP does not appear to directly disrupt KLF1 expression in K562 cells (Figure 5.15). ssKS2-10 and ssKS2-29 were the two K562 cell lines homozygous for the SNP, while ssKS2-47 was heterozygous for the SNP and the wild type allele. Figure 5.15A shows rtPCR analysis of KLF1 in these three cell lines, none of which differ significantly from the wild type K562. Although not statistically significant, KLF1 expression actually appears to have increased in these cells. The changes in KLF1 expression observed in the other cell lines are more interesting, especially since these were included as controls to assess the effect size of the SNP.

Of the three heterozygous cell lines shown in Figure 5.15B, ssKS2-3 and ssKS2-45 were almost completely depleted of KLF1 mRNA. As was described in 5.6.1.1, both of these cell lines were heterozygous for the KLF1 SNP and indel mutations, a 2bp insertion and a 1bp deletion respectively. ssKS2-4 was heterozygous for the same indel mutation as ssKS2-3, and was wild type on the other allele, meaning that the only difference between these cell lines was the presence of the KLF1 SNP on one allele. Given the clear difference in KLF1 expression levels between ssKS2-3 and ssKS2-4, this suggests that the SNP does have an effect, although why this is not observed in the absence of an indel mutation on the other allele is unclear. The same effect is observed in the ssKS2-45 cell line, where the KLF1 SNP appears to deplete KLF1 expression when an indel occurs on the other allele.

Figure 5.15C shows the KLF1 expression of the three cell lines that contained homozygous indel mutations. KKL8 had a 5bp deletion, KKL17 a 41bp deletion, and KKL11 had the same 2bp insertion that was observed in ssKS2-3 and ssKS2-4. All three of the cell lines homozygous for indel mutations around the SNP site show significantly decreased expression of KLF1, further demonstrating that sequence disruption at this locus impaired KLF1 expression.

KLF1 expression in the two PAM site disruption cell lines ssKP2-3 and ssKP2-4 are shown in Figure 5.15D. Both of these cell lines are heterozygous for the PAM disruption, and an indel mutation, a 1bp insertion in ssKP2-3 and a 12bp deletion in ssKP2-4. Reduction in KLF1 expression was observed in both of these cell lines, but it was only statistically significant in ssKP2-3. Given the strength of the reduction in the other cell lines heterozygous for the KLF1 SNP and indel mutations, this could suggest that the effect of the PAM disruption is lesser than that caused by the SNP. However this is not particularly clear, and while the difference in KLF1

expression between ssKP2-3 and ssKP2-4 is not statistically significant, the effect appears to be much stronger in ssKP2-3.

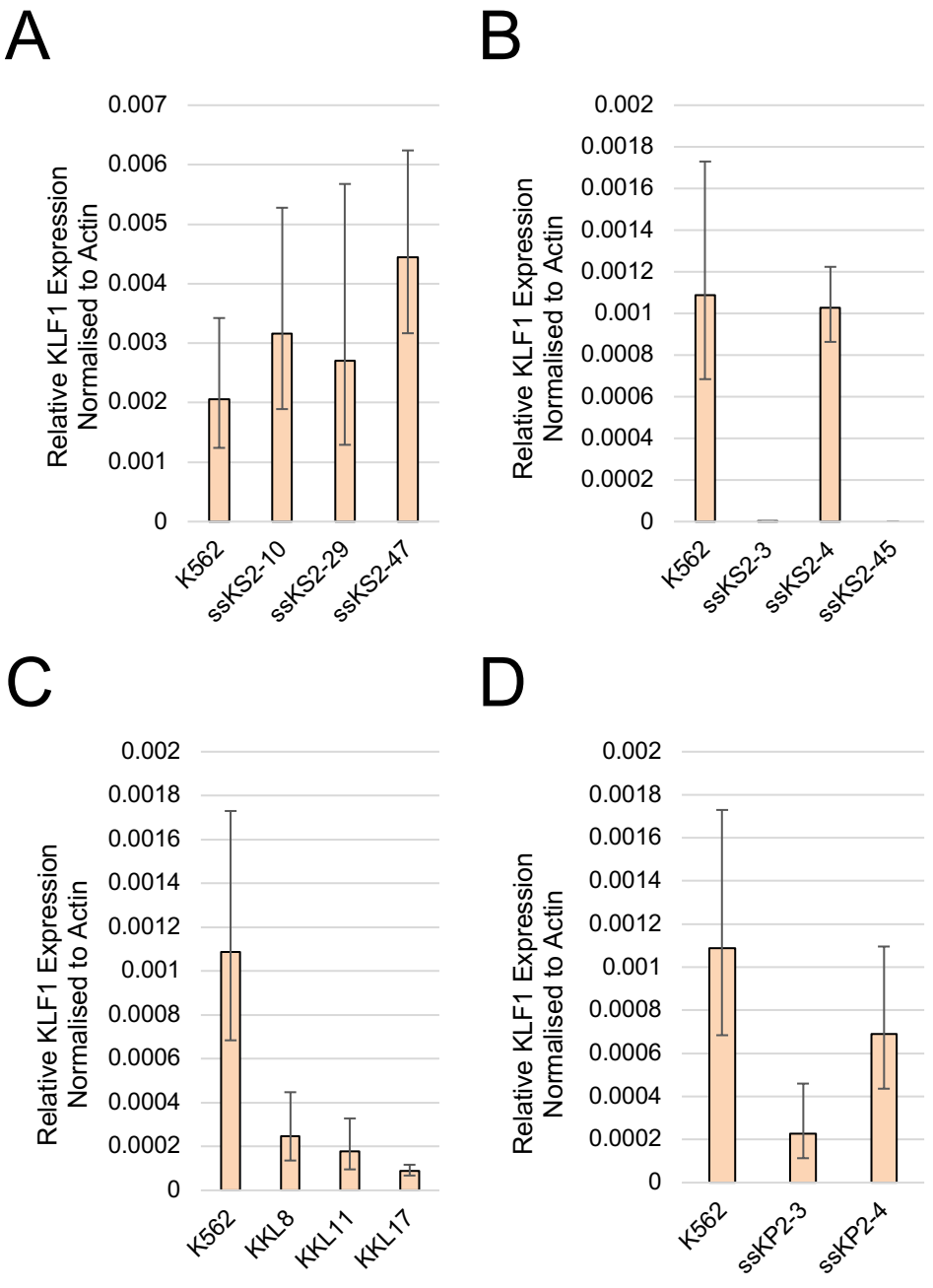


Figure 5.15: KLF1 rtPCR analysis of wt K562 and cell lines containing different KLF1 mutant genotypes. Graphs show relative expression normalised to actin β, for A – Cell lines containing the KLF1 SNP with no indel mutations. ssKS2-10 & ssKS2-29 were homozygous, ssKS2-47 was heterozygous. B – Cell lines heterozygous for indel mutations. ssKS2-3 & ssKS2-45 were also heterozygous for the KLF1 SNP. C – Cell lines containing homozygous indel mutations. D – Cell lines heterozygous for the PAM site disruption and indel mutations. Error bars indicate 95% confidence intervals, calculated from three technical replicates for each of the cell lines. KLF1 expression is not reduced in cell lines containing only the KLF1 SNPs, but is significantly reduced in cell lines containing homozygous indel mutations or heterozygous for indel mutations and the KLF1 SNP. ssKS2-3 and ssKS2-45 had extremely low KLF1 amplification, with levels the same as in the reverse transcriptase negative controls (not shown).

5.6.2.2 Globin gene expression in K562 KLF1 mutants

Expression of α-globin, β-globin and γ-globin was assayed for each of the KLF1 mutant cell lines, the results of which are shown in Figure 5.16.

For the cell lines containing the KLF1 SNP with no indel mutations, α-globin, β-globin and γ-globin expression was significantly increased in all three cell lines, but the effect size for β-globin and γ-globin expression was lesser in the heterozygous cell line ssKS2-47.

Again, a strong effect was observed in the cell lines heterozygous for the KLF1 SNP and indel mutations. ssKS2-3 and ssKS2-45, which were shown to have greatly reduced expression of KLF1, were also found to have almost completely silenced expression of α-globin and γ-globin, but strongly increased expression of β-globin. While the loss of γ-globin in these cell lines can be explained by the fact that KLF1 is required to maintain the chromatin architecture at the β-globin locus, and therefore expression is lost in its absence, the severity of the change in α-globin expression is unexpected. KLF1 is a positive regulator of α-globin gene expression, binding at the promoter, but does not have the same effect on chromatin structure as is observed at the β-globin locus, e.g. Klf1 knockout mice don't survive past early foetal stages due to severe β-thalassaemia, while α-globin expression still persists[66,68,510–512]. It has been suggested that this could be caused by the K562 cells in these clones losing their erythroid phenotype, causing the severe reduction in globin expression, contrary to what has been observed in the other K562 clones. Further investigation using KLF1 siRNA knockdowns, or using the CRISPR-Cas9 system to completely knock out KLF1 in these cells should inform as to whether the same effect is observed in response to loss of KLF1, or whether the effect observed in these cells is due to more widespread changes. Similarly, other erythroid marker genes should be tested in these cell lines, to investigate whether this effect is limited to the globin gene expression. The GATA family of transcription factors would be good candidates for this analysis, since they are key regulators of haematopoiesis, and since GATA1 is an upstream regulator of KLF1[52].

The ssKS2-4 cell line, that previously showed no change in KLF1 expression, showed increased expression of β-globin and γ-globin, but β-globin expression was approximately a quarter of that observed in ssKS2-3. Similarly, the three cell lines containing homozygous indel mutations showed increased expression of all three genes, with a very small effect size for β-globin expression.

For the PAM site disruption cell lines, ssKP2-3 which had greatly reduced KLF1 expression, also showed strongly reduced α-globin and γ-globin expression, with an increase in β-globin levels, following the same pattern as ssKS2-3 and ssKS2-45. ssKP2-4, which showed no change in KLF1 levels, also showed no change in γ-globin expression, and had slight but significant reductions in levels of α-globin and β-globin.

It is clear from these results that overall globin expression was increased in the majority of cell lines, compared to the wild type K562. Our initial hypothesis was that introduction of the KLF1 SNP, as well as other sequence disruptions around the SNP site, would reduce KLF1 expression, and in turn switch expression from β-globin to γ-globin. Since the levels of all three of the globin genes assayed were altered in most cell lines, it was decided to compare the ratio of γ-globin to β-globin, to estimate the change in proportion of globin transcripts as a result of the various genotypes being assayed (Figure 5.17).

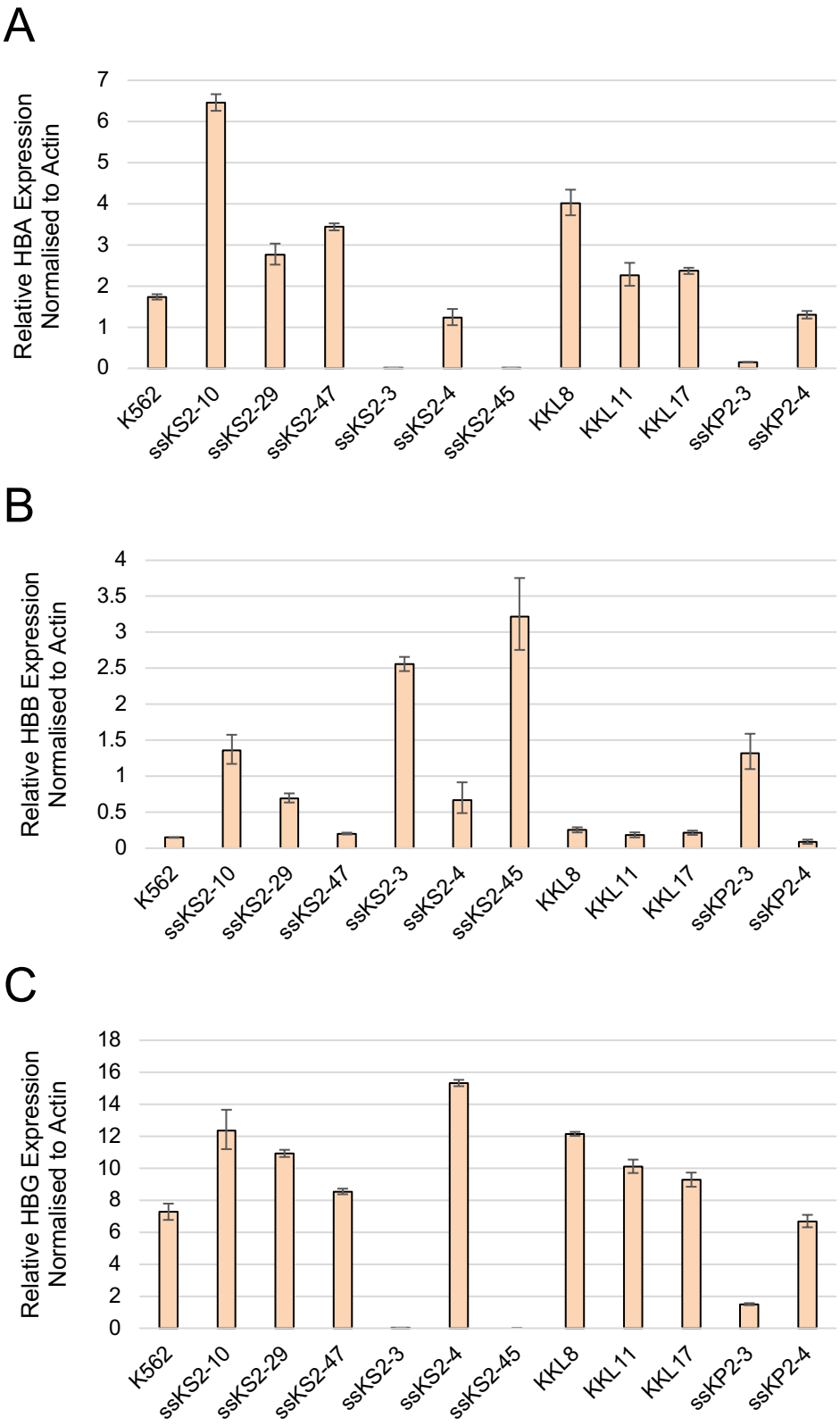


Figure 5.16: Globin rtPCR analyses of wt K562 and cell lines containing different KLF1 mutant genotypes. Graphs show relative expression of genes normalised to actin β, for A – α-globin (HBA), B – β-globin (HBB) and C – γ-globin (HBG). Error bars indicate 95% confidence intervals, calculated from three technical replicates for each of the cell lines. Total globin gene expression appears to have increased in all cell lines, apart from ssKS2-3, ssKS2-45 and ssKP2-3, where HBA and HBG decreased, and HBB increased. These three cell lines showed strong reduction in KLF1 expression in Figure 5.15.

Comparing the ratios of γ-globin to β-globin expression, it is clear that the two cell lines homozygous for the KLF1 SNP result in an increase in the proportion of β-globin transcription form the locus. Interestingly this appears to occur independently of any effect on KLF1 expression. The heterozygous cell line ssKS2-47 was not significantly changed from wild type K562 cells. Similarly, the three cell lines containing homozygous indel mutations (KKL8, KKL11 and KKL17) did not change significantly, despite the marked decrease in KLF1 expression observed in these cell lines.

As was expected, given the clear loss and gain of γ-globin and β-globin expression respectively, the γ-globin:β-globin ratio was greatly reduced in ssKS2-3, ssKS2-45 and ssKP2-3 cells. Interestingly, ssKP2-4 was the only cell line to show an increase in the proportion of γ-globin transcripts. While KLF1 expression did appear to be reduced in ssKP2-4, it was not statistically significant.

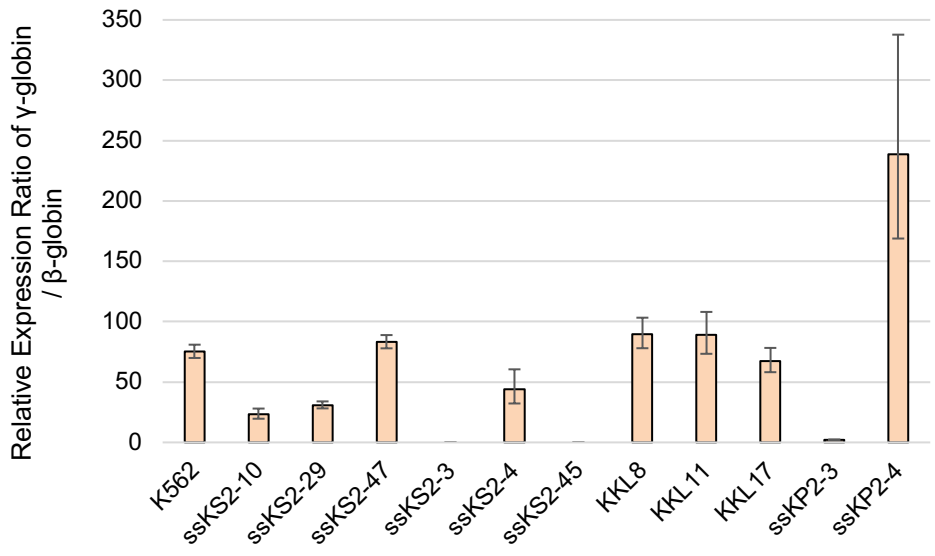


Figure 5.17: Globin rtPCR analyses of wt K562 and KLF1 mutant cell lines, showing γ-globin normalised to β-globin expression. Error bars indicate 95% confidence intervals, calculated from three technical replicates for each cell line.

Similarly to the results from the cell line containing the ASH1L SNP (KAX9), the observed effect was the opposite of what was predicted. KLF1 is known to be a positive regulator of β-globin, and loss of expression was expected to result in down regulation of β-globin and increased expression of γ-globin. However, it is possible that the observed results are an artefact of the K562 pattern of globin expression, since γ-globin is already highly expressed in these cells.

From the results generated by this study, it is therefore not possible to conclude whether the KLF1 SNP has a direct impact on KLF1 transcription. Cell lines homozygous for the SNP had

altered γ-globin:β-globin expression, but did not show any significant change in KLF1 expression, whereas compound heterozygotes with indel mutations on the alternative allele experienced a strong effect.

5.7 Summary of the CRISPR Genomic Editing Results

While the preliminary functional analyses performed on the cell lines generated by this work were inconclusive, the variants do appear to be having an effect on globin gene expression, and future work investigating these cell lines will provide more information.

More importantly, the results from this chapter demonstrate that we have improved the efficiency of our CRISPR genomic editing pipeline, and that we are able to successfully introduce specific candidate variants into cell lines *in vitro*. This pipeline will therefore be used to perform functional analyses on the nine candidate genetic modifiers identified by the SCA exome sequencing study performed in Chapter 4.

Chapter 6 Discussion

6.1 Isolation of Nucleated Erythroid Progenitors

6.1.1 *In vitro* expansion of erythroblasts is not appropriate for use in longitudinal studies

Initial attempts to isolate nucleated erythroid progenitors from small volumes of peripheral blood relied on a two phase *in vitro* culture to expand the proerythroblast population and then induce them to differentiate, with the aim of collecting cells at the late basophilic to polychromatic erythroblast stage, when cells are expressing both GPA and CD71.

In vitro cultures are commonly used in laboratories for testing the effects of drug treatments on erythroblastic populations. However, in these studies the treatment is also performed *in vitro*, whereas we intend to perform longitudinal drug treatment studies *in vivo*, and to collect blood samples at set time points throughout treatment, which would then be expanded *in vitro* for analysis.

The *in vitro* culturing of erythroid progenitors from peripheral blood was found to be extremely unreliable for blood from both healthy donors and SCA patients, with many cultures not surviving past the transition into the second phase. The low success rate of these cultures is not necessarily problematic for certain types of studies, particularly those where experiments do not rely on patient samples and are able to be repeated with minimal inconvenience. However, for the purpose of processing small volumes of blood collected from severely anaemic patients at specific time points, as was intended for this study, where the effects of HU treatment on the methylome were to be investigated, the success rate of this technique was thought to be prohibitively low.

In addition to the low culture success rates, there were also concerns regarding the influence the culture system itself could have on the methylome and transcriptome of the erythroblasts. Treating cells *in vitro* with SCF and erythropoietin has been shown to promote γ-globin expression over β-globin[513,514], mimicking the effects observed under stress erythropoiesis. The glucocorticoid receptor, which is the target for dexamethasone, is also involved in induction of stress erythropoiesis in mice[164]. This could potentially complicate elucidating the mechanism of action of drugs such as HU in a treatment study, and it may be difficult to draw conclusions from data obtained using this method.

6.1.2 $CD71^+GPA^+$ cells absent from healthy donors can be isolated from the peripheral blood of SCA patients, but lack a nucleus

A large erythropoietic population was identified in the peripheral blood of SCA (HbSS) patients that was absent in the blood samples from healthy donors, as well as less severe SCD genotypes (HbSC). These cells were initially observed as a red coloured layer present during the isolation of PBMCs, and were identified by flow cytometry to represent a $CD71^+GPA^+$ population.

Initial attempts to isolate this cell population using FACS were unsuccessful, with high rates of cell death, and a very low yield of DNA and RNA. This was improved through the use of a less stressful magnetic bead-based separation technique, depleting PBMC samples of $CD45^+$ cells before enriching for $CD71^+$ erythroblasts. The purity of the isolated population was not as high when using the magnet bead technique, since the specificity of the FACS process means that each cell is sorted on its individual fluorescence pattern, but the RNA yield was significantly increased.

Since CD71 expression *in vitro* is lost by the end of the polychromatic erythroblast stage, and GPA expression starts at the basophilic erythroblast stage (Figure 1.6), this $CD71^+GPA^+$ population was thought to represent cells between these two stages. Upon further investigation it was discovered that these cells were at a much later stage than anticipated, and actually represented enucleated reticulocytes. While this explained why the DNA yield from these cells was consistently low despite testing several different DNA extraction techniques, it was unexpected.

GPA and CD71 expression is observed on nucleated erythroblasts grown *in vitro*, and RNA-seq of erythroblasts at different stages of differentiation *in vitro* has demonstrated high CD71 and GPA expression at the late basophilic and polychromatic stages, with a marked reduction by the orthochromatic stage[131,402]. However, a cursory literature search into reticulocyte surface markers shows that CD71 is commonly used as a marker for immature reticulocytes, and that these cells are increased in SCA patients, likely as a result of stress erythropoiesis[515–518].

This suggests that there is discrepancy in surface expression patterns of GPA and CD71 on maturing erythroid progenitors grown *in vitro* and *in vivo*, and highlights the concerns mentioned in 6.1.1 about how the culture may alter erythroblastic development. This is supported by the fact that reticulocytes plated in an *in vitro* culture rapidly lose expression of CD71[519].

Even with the $CD71^+GPA^+$ cells *in vivo* being more mature than anticipated, it is surprising that no DNA was extracted from this population, given that Walker *et al.* used the same isolation technique to investigate SCA patients before and after treatment with HU, and were able to extract DNA from these reticulocyte populations[245]. Due to the high numbers of cells isolated in that study, and the small amount of DNA required for the locus specific bisulphite sequencing that was performed, it is perhaps possible that enough DNA was extracted from a very small proportion of nucleated erythrocytes. Alternatively, with a purity of >90% after CD71 enrichment, this DNA could be provided by contamination of other haematopoietic lineages in the sample[245].

Overall, these results demonstrated that while the cell surface expression markers chosen to select for erythroid progenitors at a specific stage were correct for use for the *in vitro* cultured cells, they are not appropriate for the isolation of human erythroid progenitors differentiating *in vivo*.

6.1.3 Isolation of early stage progenitors from the peripheral blood of SCA patients

Due to the unexpected discovery that the $CD71^+GPA^+$ cells had already undergone enucleation, it was decided to instead attempt to isolate erythroid progenitors at a much earlier stage of development. CD34 is an early stage marker of haematopoietic cells that is lost before reaching the proerythroblast stage[174]. $CD34^+$ cells were successfully isolated, but the total cell numbers obtained from 9ml of blood were too low for efficient extraction of DNA and RNA.

Interestingly, two populations of $CD34^+$ cells were identified, co-expressing either the early stage marker CD45 or the late stage marker GPA. This was unexpected given that GPA is normally expressed much further down the developmental pathway, after CD34 expression is lost. These $CD34^+GPA^+$ cells have been documented previously, and are believed to occur as a result of stress erythropoiesis, representing a population of progenitors undergoing accelerated differentiation, where CD34 downregulation does not occur[520].

Rather than enriching for markers of early stage differentiation, depletion of markers of late stage differentiation was found to be more effective. GPA^-CD71^+ cells were successfully isolated, almost all of which were co-expressed with either high or low levels of CD45 expression, presumably representing the stages of erythroid development prior to the loss of CD45. The purity of the $GPA^-CD45^+CD71^+$ population was lower than that of the CD34+ enriched samples, with approximately 83% purity compared to 97%, however the total number of cells was much greater, and DNA was successfully extracted from these cells.

Using the GPA depletion followed by CD71 enrichment, we were able to obtain suitable quantities of DNA from nucleated erythroid progenitor cells directly extracted from the peripheral blood of SCA patients, including one patient undergoing HU therapy. Although the reduction in purity may affect the sensitivity of downstream analyses, making it more difficult to detect smaller changes, we believe that this technique could be used in longitudinal studies and provide valuable insight into the mechanism of action of treatments such as HU in erythroid progenitor cells *in vivo*.

6.2 Identification of Candidate Genetic Modifiers of SCA by Whole Exome Sequencing

In addition to epigenetic factors, genetic polymorphisms are known to influence the severity of the SCA phenotype, and we also set out to identify novel genetic modifiers by conducting a WES study.

WES is a powerful tool for the identification of novel genetic variants either modifying or causative of disease phenotypes. A large number of variations from the reference genome are annotated within each individual, the majority of which occur in non-coding regions, since these are more tolerant of variation without influencing gene function. This is the case despite the fact that the commercially available exome capture kits specifically target protein-coding genes and ncRNA, and of the 2,798,560 variants that were annotated from 19 mild SCA exomes in this study, 2,662,432 occurred outside of ncRNA and non-protein coding sequences.

6.2.1 Exome Variant Filtering Pipeline

The remaining 136,128 candidate variants were pared down to 3,159 by the application of a series of filtering criteria, designed to remove variants that are unlikely to modify the SCA phenotype. These steps included the removal of genes commonly mutated in exome sequencing studies, as well as genes that were identified as not being expressed in haematopoietic tissues, as identified in publicly available gene expression data. The use of this pipeline greatly improved our ability to identify potential candidate variants, and removed many of the variants identified by the other analyses conducted in this study, that were thought to be false positive results, such as the variant in FSIP2 which was significantly associated with the mild SCA patient group, but is only expressed in spermatocytes.

An imbalance in the representation of ncRNA in the final candidate variant list compared to the initial coding variant list was observed, making up 2,563 of the 3,159 variants. The fact that so many variants in ncRNAs remain post filtering, despite making up <1% of initial variants suggest that the ncRNA are not efficiently targeted by the filtering criteria. The proportion of ncRNA first noticeably increases after the filtering out of variants from the severe SCA exomes, rising from 15.2% to 37.9% and 42.9%. This may be because ncRNA are more tolerant of genetic variation. If higher levels of ncRNA variation are observed between the US and the UK SCA populations, then that might explain why filtering using the severe US datasets is less effective than it is for

variants in the protein coding regions. Similarly, due to the small size of the UK severe SCA patient group, fewer of the specific ncRNA variants may be present in this control population.

During the variant filtering process, two parallel candidate variant lists were generated, with the US severe SCA exome group consisting of either the SWiTCH exomes only, or both the SWiTCH and TWiTCH exomes. This was undertaken due to concerns about the definition of the 'severe' phenotype group, with SWiTCH participants having already suffered a stroke, whereas TWiTCH participants were identified as being at risk for stroke. As such, the SWiTCH participants more closely matched the phenotype of the severe patient group recruited in the UK. The benefit of including the TWiTCH exomes was that it doubled the size of the severe cohort from 137 to 276 exomes, and reduced the final number of candidate variants from 3,159 to 2,597.

We believe that the additional filtering power provided by including the TWiTCH exome data was not worth sacrificing the strict clinical definition of the severe cohort, and two of the most interesting candidate variants identified were excluded when combined with the TWiTCH data. Even though one of the SCA patients from the UK severe group had not experienced a stroke, they were included based on the severity of the other symptoms at a young age, and based on the opinions of our clinical collaborators, whereas no clinical information was available for the individual TWiTCH participants.

We demonstrated that using the variant filtering pipeline developed during this project we were able to detect seven putative genetic modifiers of the SCA phenotype, with biologically plausible mechanisms to influence the pathology of the disease. The major limitation of this analysis was highlighted by the fact that we were unable to directly detect the β^0-thalassaemia mutation and the novel KLF1 variant due to the fact that they were only present in one patient, and we filtered out single occurrence variants. The small sample size of the mild SCA group is therefore the limiting factor for detecting rare variants. Both of these variants were identified by specifically investigating variants in known modifier genes before the filtering of single occurrences.

The variant filtering pipeline was also used to generate a list of variants to identify commonly mutated genes in the mild SCA patient group. The utility of this analysis was limited, and highlighted the ability of ncRNA and some coding genes (such as MUC22) to tolerate large genetic variation. Of the protein coding genes containing a small number of variants that affected a large proportion of samples, none presented with a biologically plausible mechanism by which to affect SCA pathology. Since we were investigating modifier variants, and not

variants causative of a disease phenotype, we decided not to select for rare variants during the filtering process. The inclusion of common variants generates a high proportion of false positive 'hits' in the analysed genes, and is likely to be one of the reasons why no plausible candidates were identified by this analysis.

6.2.2 Differing genetic ancestry between the UK and the US SCA groups

In the exome sequencing study, all of the mild patient group were recruited and sequenced from the cohort at King's College Hospital in the UK, whereas the severe patient group consisted of 132 patients recruited as part of the SWiTCH trial from centres across the USA, as well as five additional patients from King's College Hospital. This means that as well as the two test groups being defined by severity of disease, they are also separated geographically. There is large genetic heterogeneity between different African populations[521,522], which can be observed in the distribution of the different sickle haplotypes (Figure 1.8). Increased heterogeneity would be expected between populations of African ancestry living in the USA and those in the UK, with variation arising as the result of the different migration events of different African sub-populations, as well as due to admixing with other populations in either the USA or the UK.

The problem of genetic heterogeneity between populations within genetic association studies is common, and can be overcome by allelic clustering, stratifying test groups into smaller groups based on haplotype[522]. However, due to the small sample size of the mild group, stratification by haplotype was not performed, since any reduction in false positive association would be offset further by a reduction in sensitivity to detect associations if the mild group were segregated into even smaller subgroups.

The way to overcome this problem is to vastly increase the number of patients in the exome cohort. Future work would benefit from expansion of the sample size for both the severe and mild patients within the UK, and also the inclusion of mild SCA patients from the USA. Aside from simply increasing the sample number, this would allow a balanced stratification of genetic association based on haplotype.

With the observed association of variants based on genetic ancestry rather than by phenotypic severity, it was decided not to perform statistical testing on the candidate variants identified by the variant filtering pipeline. In this analysis, the severe exomes were used to generate a negative filtering list, where it was assumed that presence of a variant in the severe group meant that it was not protective of the severe SCA phenotype, but absence of a variant from the

severe group was not assumed to represent positive association with the mild group. Due to the design of this analysis, no false positive candidates were introduced as a result of the differences in genetic ancestry. While it would be possible to test the allelic imbalance of the candidate variants between the two groups using a Fisher's Exact Test, the results of this test would be misleading, since as well as having artificially enriched for variants absent from the severe group, the statistical test would be biased by differences in genetic ancestry between the two groups.

6.2.3 Statistical testing of association of variants with SCA phenotype groups

Fisher's Exact Tests for association of variants with either the 19 mild patients or the 5 severe patients recruited from King's College Hospital only reached statistical significance for seven variants, all of which occurred in non-coding regions. This was likely due to the small sample size of each group. By including the exomes from the SWiTCH clinical trial in the severe group, 2,442 variants reached statistical significance (Appendix 8). However, due to the clear bias generated by differences in genetic ancestry between the two groups, no candidate variants identified by this analysis were considered robust enough to warrant further investigation.

For the statistical analyses investigating association of variants with either the SWiTCH exomes or the HUSTLE exomes, all patients were recruited from SCA populations within the USA, and so the bias introduced by differences in genetic ancestry is minimised. This analysis still considered the SWiTCH group to represent the severe SCA phenotype, but instead of being compared to a small mild group of 19 patients, the HUSTLE group was considered to be representative of the general SCA population, and had a much larger sample size of 140 patients. Using this analysis 236 coding variants reached statistical significance (Appendix 9), and we were able to identify candidate variants in genes with biologically plausible mechanisms through which to influence the phenotypic severity of SCA.

6.2.4 Candidate Modifier Genes and Variants

Of the candidate variants identified by the variant filtering pipeline and the comparison of SWITCH and HUSTLE exomes, it was decided that seven warrant further investigation. The single occurrence variants in KLF1 and HBQ1 were also included, due to the established roles that KLF1 and the globin genes have in affecting the SCA phenotype. These nine variants fall into four distinct mechanistic groups: Nitric Oxide Signalling, Haematopoietic Regulation,

Altered Globin Gene Expression and Recovery from Ischaemic Injury. Future work in the laboratory will use CRISPR genomic editing to investigate the effects of these SNPs on gene function *in vitro*.

6.2.4.1 Nitric Oxide Signalling: NMRAL1

The heterozygous stopgain variant in NMRAL1 was observed in two mild SCA patients. NMRAL1 operates in a negative feedback loop with argininosuccinate synthetase, which is rate limiting for NO production. Increased expression of argininosuccinate synthetase upregulates expression of NMRAL1, which in turn inhibits argininosuccinate synthetase activity[425]. This presents a mechanism by which haploinsufficiency for NMRAL1 could increase argininosuccinate activity, in turn increasing NO production and resulting in a vasodilatory effect, which could reduce the frequency and severity of vaso-occlusive events. By this mechanism the heterozygous stopgain in NMRAL1 could achieve the same outcome targeted by arginine therapy[262].

RNAi mediated knockdown of NMRAL1 in HEK293T cells has previously demonstrated an increase in NO production[425,523]. CRISPR genomic editing could be used introduce this stopgain into a cell line expressing NMRAL1, allowing confirmation that the variant reduces NMRAL1 production, and that this results in an increase in NO levels. Since the NMRAL1 variant was heterozygous in the mild SCA patients, ideally this mutation would also be introduced into the cell line in a heterozygous manner, to confirm that there is no compensation by the other allele.

The role of NO in the pathophysiology of SCA is currently disputed, so while reconstituting the NMRAL1 mutation in a cell line may demonstrate its effect on NO levels, it would not validate its effect as a modifier of SCA[23]. To further investigate this, a SCA mouse model may be used, such as the one available from the Jackson Laboratory (www.jax.org - strain: 003342)[524].

6.2.4.2 Haematopoietic Regulation: IGFBP2, FLT3, ETS2, MALAT1 & BAG1

Variants were identified in five genes that influence the survival and proliferation potential of haematopoietic populations. The variants in IGFBP2, FLT3, ETS2 and MALAT1 were identified by the analysis using the variant filtering, and would be expected to ameliorate the phenotype severity of SCA. The variant in BAG1 was identified by the statistical analysis of variants in the SWiTCH and HUSTLE exome groups, and was significantly increased in the SWiTCH group,

representing patients that had experienced a stroke at a young age, and so is predicted to exacerbate the severity of symptoms.

There are two main mechanisms by which altered regulation of erythropoiesis could affect symptomatic severity of SCA. The first is by altering normal steady state erythropoiesis, where increased erythropoietic potential would result in an increased number of sickling erythrocytes as well as increasing the concentration of cells in the blood, both of which could lead to an increase in vaso-occlusive events. Therefore variants that promote steady state erythropoiesis would be expected to increase severity, whereas those that reduce steady state erythropoiesis would be expected to ameliorate severity.

The second mechanism is by altering stress erythropoiesis, where higher production of F-cells reduces the amount of sickling observed, and results in a reduction in vaso-occlusive events[143]. Therefore variants that promote stress erythropoiesis would be expected to ameliorate disease severity, whereas those that repress stress erythropoiesis would be expected to exacerbate phenotype severity.

These two processes do not exist independently, and it is likely that variants affecting one will have an impact on the other. For example, a variant reducing steady state erythropoiesis will likely trigger an increase in stress erythropoiesis in response to hypoxia.

While IGFBP2 also has an intracellular function, both IGFBP2 and FLT3 are targeted for secretion, and function as extra-cellular signalling molecules[525]. IGFBP2 promotes survival and expansion of haematopoietic stem cells in the bone marrow, and FLT3 is membrane bound tyrosine receptor kinase that stimulates expansion and migration from bone marrow into the peripheral blood[117,429,437]. The variants in IGFBP2 and FLT3 both occur in the N-terminal signalling peptide, required for secretion. Since the signal peptides are cleaved, these variants are unlikely to have an effect on function, but could interfere with secretion. Introducing these variants into cell lines will allow investigation into how localisation of the proteins is affected, and could be assayed using fluorescently tagged antibodies, either using fluorescent microscopy or in the case of the membrane protein FLT3, flow cytometry.

ETS2 is a transcription factor important for lineage determination in a number of tissues at different stages throughout development, and in Megakaryocyte-Erythroid Progenitors its expression is thought to promote the megakaryocyte pathway over erythrocyte development[526]. When overexpressed in K562 cells, ETS2 increases expression of TAL1, but reduces expression of the erythroid specific transcription factor KLF1, and subsequently β-globin[443]. To

investigate the effect of the ETS2 variant *in vitro*, plasmids constitutively expressing either ETS2 or the ETS2 mutant could be transfected into K562 cells, and expression of both KLF1 and the globin genes assayed by rtPCR. Given that loss of function of ETS2 would likely result in an increase in erythrocyte production, this variant would be expected to result in a gain of function in order to ameliorate the disease phenotype, perhaps enhancing the signal for cells to differentiate into megakaryocytes rather than erythrocytes. Alternatively, if a loss of function is observed, then it would be assumed that loss of ETS2 function allows greater erythrocyte expansion under conditions of stress erythropoiesis.

MALAT1 is a ncRNA, and is thought to influence haematopoietic development through repression of B-MYB, a positive regulator of haematopoietic potential[444,446,447]. The effect of the MALAT1 variant after introduction into a cell line *in vitro* could be assayed by detecting B-MYB expression levels, including rt-PCR of differently spliced isoforms, since MALAT1 is known to affect serine/arginine splicing factors[445].

The BAG1 variant occurs in an arginine rich domain, which is only present in BAG1L and BAG1S, two of the four alternative splicing isoforms. BAG1L is the only isoform that is recruited to the nucleus, and it has been shown that BAG1L mutants lacking residues 17-50 show reduced nuclear localisation[527]. The variant in BAG1 changes glycine at position 45 to arginine, and since densely packed positively charged arginine and lysine residues are important for nuclear import, it's possible that this variant could increase efficiency of nuclear localisation[528].

BAG1 markedly increases the anti-apoptotic activity of BCL2, and if nuclear localisation is increased as a result of this variant, then it is reasonable to expect that this anti-apoptotic effect would be increased as well[529,530]. Apoptosis plays an important role in regulating the rate of haematopoiesis, and limits the number of progenitor cells that reach full maturity, BCL2 plays an important role in this process, and overexpression in mice more than doubles the number of HSC in the bone marrow[531]. Interestingly the balance of BCL2 to BCL-X appears to influence lineage commitment, with higher BCL2 promoting the granulocyte lineages, and BLC-X promoting the erythroid lineage. This affect is thought to be mediated by RAF-1 expression levels[532].

Introduction of this variant into a cell line by CRISPR genomic editing will allow investigation into the effect that the variant has on intracellular localisation of BAG1. It would also be interesting to investigate whether the variant has an effect on survival, particularly when cultured in the

presence of pro-apoptotic factors. Quantification of RAF-1 expression in these cell lines could determine whether or not the BAG1 variant influences BCL2 directed lineage commitment.

6.2.4.3 Altered Globin Gene Expression: KLF1 & HBQ1

Variants were identified in two genes affecting globin gene expression and function. Variants in KLF1 and HBQ1 were exclusive to the mild SCA patient group, but each only occurred in a single patient, and so the presence of these variants in other mild SCA patients should be confirmed before laboratory investigation.

The effect that altered expression from the globin gene loci can have on the phenotype severity of SCA is well established, and is described in 1.6. The main principle behind this mechanism is that polymerisation of haemoglobin only seems to occur in the HbS ($\alpha_2\beta^S_2$) tetramer, and increasing the abundance of other globins introduces competition for tetramer formation between the different subunits, diluting the intracellular concentration of HbS and in turn reducing rates of erythrocyte sickling.

The KLF1 variant is a novel SNP identified in one of the mild SCA patients, and was not observed in any of the other SCA exomes or in dbSNP[533]. The SNP results in an arginine replacing the histidine at position 329, which is involved in coordination of the Zn^{2+} ion in one of the three zinc finger domains[534]. The variant was heterozygous, and so could exhibit an effect similar to haploinsufficiency for KLF1, which is known to increase γ-globin and ameliorate the SCA phenotype[535]. The patient in which this was observed also had the highest HbF% of all the patients investigated. This variant could be investigated by introducing the SNP into a cell line using CRISPR genomic editing, and assaying for changes in the γ-globin:β-globin expression ratio.

HBQ1 encodes θ-globin, and due to the fact that expression from the gene is minimal, it seems unlikely that any variant affecting gene function would have an impact on SCA phenotype. However, the variant occurs approximately 90bp downstream from the transcription start site, and falls within binding sites for multiple transcription factors, including CTCF and 7BTB7A (as annotated by data from the ENCODE Consortium[472,473]), and could potentially disrupt regulation. Introducing this variant into a cell line using CRISPR will allow detection of any upregulation of HBQ1 expression, if this is not observed then it can be assumed that the SNP does not influence disease phenotype. An alternative strategy would be to directly assay the blood of this

patient, and to determine if abnormal levels of θ-globin are observed at the protein level in erythrocytes.

6.2.4.4 Recovery from Ischaemic Injury: MYDGF

Vaso-occlusive events are responsible for the most severe symptoms associated with SCA. While factors affecting the ischaemic response would not be expected to prevent these vaso-occlusive events from occurring, the long term effects of these events could be ameliorated by an efficient response, reducing levels of cell death and tissue damage experienced by minimising the amount of time that affected areas remain starved of oxygen. Many SCA patients experience silent vaso-occlusive events in the brain, that occur in minor blood vessels and present with minimal phenotype, and are typically not diagnosed in the clinic [536–538]. It is perhaps possible that impaired ischaemic repair pathways could increase the severity of these events, increasing cell death and the resultant inflammatory response to the extent where they become observable and present as an overt clinical stroke.

The variant observed in MYDGF was significantly increased in the SWiTCH study group compared to the HUSTLE group, and so would be expected to increase severity of the SCA phenotype. MYDGF (Myeloid Derived Growth Factor) is a growth factor initially identified by its function in myocardial cell growth and survival following myocardial infarction, which stimulates its secretion by monocytes and macrophages[466]. Mice deficient for MYDGF have no discernible phenotype until myocardial infarction is induced, and then demonstrate increased scarring and reduced angiogenesis at the infarction site, as well as reduced systolic and diastolic function[466]. MYDGF mediated protection against cell death in mice was found to be dependent on the PI3K signalling pathway, and so the effect of the variant on MYDGF function could be investigated by introducing this SNP to a cell line *in vitro* using CRISPR, and assaying for PI3K activation by western blot for phosphorylation of PI3K as well as downstream targets BAD and BAX, comparing the results to those observed in wild type cells, and cells treated with PI3K inhibitors[466].

6.3 Generation of Mutant K562 Cell Lines and Preliminary Testing of Variants in KLF1 and ASH1L using CRIPSR-Cas9 Genomic Editing

We successfully introduced both the ASH1L and KLF1 SNPs into K562 cell lines *in vitro*, using CRISPR-Cas9 genomic editing.

gRNA directed introduction of DSBs at the target sites by Cas9 was reliable, but survival rates of single cell K562 cultures were low, and template incorporation by the endogenous DNA repair pathways was inefficient. As a result this process took significantly longer than was anticipated, relying heavily on the stochastic chance of correct template incorporation on both alleles. This highlights some of the current limitations of CRISPR for genomic editing, firstly in that the endogenous repair machinery strongly favours the NHEJ pathway, and secondly that until correct template incorporation is much more efficient, cell populations must be expanded from a single cell in order to generate a genetically homogenous population.

However, through optimisation of the CRISPR-Cas9 protocol, we increased efficiency of template uptake, and demonstrated that we are able to generate the desired mutations in cell lines *in vitro*. This CRISPR genomic editing pipeline will therefore be used in future to validate the candidate genetic modifiers identified by the exome sequencing study.

While the work in this thesis was focussed on introducing variants into K562 cells, this technique could also be used in other cell lines. It would be particularly interesting to introduce variants of interest into HUDEP-2 cells, a recently established human erythroid progenitor cell line that can be induced to develop into terminally differentiated erythrocytes[489].

6.3.1 Low transfection efficiency and low survival rates of single cell K562 cultures

Transfection efficiency was very low, and even nucleofection, which proved to be the most effective technique only resulted in 15% of cells taking up the 9.8kb Cas9 plasmid. It is not clear why this was the case, but ultimately it did not affect the results since a maximum of 288 cells from each transfection experiment were isolated by FACS.

In order to improve the transfection efficiency in future experiments, more modern nucleofection protocols should be investigated. The Nucleofector™ 2b, which was used in this project, was first introduced in 2001, and uses a cuvette based system, with the current applied to a 500µl cell suspension. More modern nucleofection machines, such as the Neon® Transfection System (first introduced in 2006), and the 4D-Nucleofector™ (first introduced in 2010), work

with much smaller volumes of cell suspension, and increase both cell viability and transfection efficiency[539]. Comparing the Nucleofector™ protocols for K562 cell nucleofection, available on the Lonza Cell and Transfection Database, it appears that using the 4D-Nucleofector™ rather than the older Nucleofector™ 2b, increases efficiency from approximately 77% to >90%, and they report >95% cell viability[540]. It is worth noting that the plasmids used to test these protocols were approximately 2.5kb, and are much smaller than the plasmids used in our CRISPR-Cas9 system. While the low transfection efficiency was not a major issue for the aims of this study, as mentioned above, it is possible that the additional stress associated with this nucleofection technique had an adverse effect on cell survival in the downstream stages of the process.

After the transfection and FACS process, very few cell lines (170/1920) survived the single cell culture stage. It is believed that this was due to the stress involved in these processes, which could perhaps be avoided in future through the use of conditioned medium during the single cell culture phase, and by replacing the FACS process with serial dilutions to obtain single cell cultures instead[398]. However, while reducing the mechanical stress associated with FACS, the serial dilutions would not allow for selection based on expression of the GFP reporter, and also would not guarantee a single cell culture. Since the process assumes that the cells are evenly distributed throughout the medium, whereas in practice some aliquots would contain more than one cell, and others would contain none. Given the low transfection efficiency, this would mean that approximately only 15% of the cultures expanded would have taken up the plasmid, and any cells demonstrating two different genotypes could not be assumed to be heterozygous, but could represent two distinct cell populations within the same culture.

If the transfection efficiency can be increased through the use of more modern nucleofection techniques, then serial dilution should be considered to increase cell viability. Even with the low number of clones that survived in this study, the genotype screening process was the rate limiting step, with PCR amplification and Sanger sequencing being carried out for each clone. With increased viability of single cell cultures, in the future it would be more efficient to introduce a pre-screening step to assess the genotypes of the transfected population prior to separation into single cell cultures, using a technique such as TIDE[541]. This would provide an accurate estimate as to whether or not the genome-editing process has generated the genotypes of interest for each transfection, and therefore whether the experiment is likely to yield the desired cell lines. Recognising extremely low success rates at this stage would avoid spending weeks expanding and sequencing unsuccessful single cell cultures, when the time could be better

spent optimising the CRISPR/Cas9 process. Additionally, the process of screening the individual clones could be streamlined, by initially performing restriction enzyme based analyses such as RFLP of RGEN on all clones, which would be faster than Sanger sequencing[542]. Individual clones that indicate a disruption of the recognition site could then be validated by sequencing.

The survival rates of the cell lines varied depending on whether they were transfected with the plasmid only, or co-transfected with either siRNA or ssODN. Counter-intuitively, the cells transfected with only the plasmid had the lowest survival rates, whereas cells co-transfected with siRNA targeting Ligase IV or Ku70 had the highest rates of survival, despite the fact that both of these genes have been shown to have anti-apoptotic effects[543,544]. Due to the limitations on transfection volume, half as much plasmid was used when co-transfected with either siRNA or ssODN which may explain the observed effect, since K562 have previously been reported to demonstrate reduced viability after transfection with increasing plasmid concentrations[545]. Therefore lowering the plasmid concentration, while having an adverse effect on transfection efficiency, could increase survival in these cells.

6.3.2 siRNA Knockdown of the NHEJ Pathway

siRNA knockdown was performed on components of the NHEJ pathway in order to promote the HDR pathway for repair of the DSBs introduced by Cas9. Rates of HDR were still very low in these experiments, but was significantly increased in the knockdowns of Ligase IV, which generated three homozygous cell lines. However, none of these cell lines had incorporated the template sequence, and this was only achieved in the homozygous cell line generated by the Ku70 knockdown experiments. Despite the statistically significant increase in HDR, the rate at which it occurred was still very low, and due to the small sample size the test is quite unreliable, since introducing just one homozygous variant into the plasmid only group would be enough to completely remove any significance.

The fact that HDR was found to be significantly increased in the Ligase IV knockdown but not the Ku70 knockdown could be due to an alternative NHEJ pathway that has recently been suggested. Fattah *et al.* found that NHEJ activity was lost in the absence of Ligase IV, but not Ku70, and that loss of Ku70 was sufficient to rescue NHEJ activity in the absence of Ligase IV[501]. They suggested that an alternative pathway compensates for the loss of the canonical

NHEJ pathway, but that it is unable to function in the presence of Ku70, which binds to the DSB, and could be out-competing this alternative pathway[501].

It is also possible that the small increase in HDR that was observed is the result of the inefficiencies of siRNA knock down, which only degraded approximately 50% of the Ligase IV transcripts. Knocking out as well as knocking down individual components of the NHEJ pathway has been investigated in silkworms, and demonstrated that full gene knockouts had a stronger effect on HDR frequency than siRNA knockdowns[354]. However, the advantage of using the siRNA knockdown technique is that its effect is transient, allowing the cells to recover after a short period of time. This allows us to reduce activity of the NHEJ pathway during the post transfection window, but does not lead to the long term genomic instability observed in stable knockouts, which would reduce the viability and reliability of our cell lines in the long term.

6.3.3 Increasing Template Uptake

ssODN templates were co-transfected alongside the plasmids in order to increase the intracellular copy number of the template sequence, as it was thought that this could be a limiting factor for template uptake. In the case of the plasmid only transfections, the template was present at a 1:1 ratio with the plasmid, and was therefore limited by the low transfection efficiencies observed.

Co-transfecting with ssODN templates did not affect the rate of HDR, which remained at similar levels to those observed previously. However, the rate of template uptake was greatly increased, with 37.8% of cells either the heterozygous or homozygous for the SNP of interest, compared to 14.7% of the cells co-transfected with siRNA. Co-transfection with the ssODN templates generated two cell lines that were homozygous for the SNP of interest.

Alternative techniques are available to increase template uptake, but were not tested in this project. These include the use of paired gRNA, which can be used to excise a large genomic fragment, forcing the cell to use the artificial template for repair. This is recommended for genomic editing of larger regions[546]. Using ssODNs with phosphorothioate modified ends can also greatly increase template incorporation, likely due to increased stability and protection from degradation[547]. Since the ssODNs used in this project were not protected from degradation by phosphorothioate modification, it is likely that the amount of intracellular ssODN template is reduced rapidly following nucleofection. Coupled with the fact that the gRNA and Cas9 were provided by a plasmid, and therefore must be generated within the cell after nucleofection, it is

possible that by the time Cas9 induced DSBs are introduced, the levels of template are already very low. This could be overcome by the direct nucleofection of Cas9/gRNA ribonucleoprotein complexes (RNPs) alongside the phosphorothioate modified ssODNs[548,549]. This might also address the low transfection efficiency observed with the large plasmids in this study.

Interestingly, It has recently been observed that gRNA targeting the strand antisense to transcription result in increased rates of HDR compared to those targeting the transcribed strand[340]. Similarly, it has been shown that Cas9 releases the strand non-complementary to the gRNA earlier, and therefore ssODN templates complimentary to this strand (i.e. complementary to the gRNA), can bind at an earlier stage and increase rates of HDR[470]. In the same study, it was also demonstrated that asymmetric ssODN design can increase HDR efficiency, with homology arms of 36bp and 91bp on the PAM distal and proximal sides of the DSB respectively[550].

6.3.4 Is the ASH1L SNP likely to cause β-Thalassaemia in patients?

It was hypothesised that introduction of the ASH1L SNP would reduce expression of β-globin in K562 cells, since the SNP was initially identified as a candidate causative mutation for β-thalassaemia. However, the results presented here suggest that the SNP had the opposite effect, increasing expression of both α-globin and β-globin, whilst reducing expression of γ-globin in the KAX9 cell line. If this SNP has the same effect *in vivo*, then it is unlikely to be causative of β-thalassaemia.

However, it is possible that effect of the ASH1L SNP *in vivo* is less specific to the reduction of β-globin expression, but instead has a more general effect disrupting regulation at the β-globin locus. Globin expression in K562 cells is more similar to that found at earlier stages of development, before the foetal globin to adult globin switch, with high levels of γ-globin expression, and low levels of β-globin. A possible model to explain the results observed could be that loss of targeted H3K4 tri-methylation at the γ-globin promoters disrupted the strength of the regulatory control, allowing increased transcription from the other promoters in the locus. This would account for the difference in our results compared to the effects previously observed in ASH1L shRNA knock downs carried out in human erythroid progenitor cells *in vitro*[476].

This model is further supported by the fact that ASH1L occupies the LCR in mice[478], and is likely recruited to the promoter of the active globin gene by chromatin looping, meaning that

disruption of function would affect transcription of the globin gene that is being upregulated at each developmental stage, and would not necessarily be specific to either β-globin or γ-globin.

It has also been found in K562 cells that H3K4 tri-methylation is associated with the active γ-globin promoters, while H3K4 at the β-globin promoter is predominately mono-methylated, which also supports this explanation of the results[551].

To further understand the effects the ASH1L SNP has on β-globin expression in K562 cells, additional analyses will be carried out on the KAX9 cell line by the laboratory in the future, specifically using Chromatin Immunoprecipitation (ChIP) to assess any changes in ASH1L occupancy at the locus, as well as changes in H3K4 methylation patterns. Since little is known about the function of the serine rich domain in which the SNP is situated, it is not known whether to expect to observe a loss of recruitment of ASH1L to the locus, or whether occupancy will remain unchanged, but with catalytic activity impaired.

6.3.5 Is the KLF1 SNP likely to affect HbF levels in patients?

The analysis of the KLF1 mutant cell lines that were generated did not provide a clear answer to whether or not the KLF1 SNP affects KLF1 expression. Cell lines either homozygous or heterozygous for the SNP showed no significant change in KLF1 expression, which suggests that the mutation is unlikely to be influencing phenotype. However, when an insertion or deletion occurred on the other allele, a strong reduction in KLF1 expression was observed, but only in the presence of the SNP. Given both of these results, it seems that the KLF1 SNP is not sufficient to affect KLF1 expression on its own, but only when the alternative allele is impaired.

KLF1 expression was also reduced in the case of homozygous insertions or deletions, suggesting that these were sufficient to reduce KLF1 expression. However, in the cell line with the wild type allele and the same insertion on the other allele, no change in KLF1 expression was observed, which could be explained by a compensatory mechanism, active on the wild type alleles but ineffective on alleles with the KLF1 SNP. A possible candidate for a feedback mechanism could be ZBTB7A, the binding site of which encompasses the site of the KLF1 SNP, and is itself positively regulated by KLF1[82,473,552]. While a model such as this would explain the results in this study, no such mechanism has been documented elsewhere. Patients with KLF1 haploinsufficiency caused by large scale deletion encompassing the entire KLF1 gene present

with high HbF expression, demonstrating that reduced KLF1 expression is not compensated for in these patients[535].

While the cell lines that were homozygous for insertions or deletions at the SNP site showed demonstrated a reduction in KLF1 expression, this did not correlate with changes in γ-globin:β:globin ratio. Interestingly, KLF1 expression was reduced to approximately the same extent in the cell line with heterozygous insertion and PAM disruption (ssKP2-3), and in this cell line a strong reduction in γ-globin:β-globin was observed, similar to that observed in the cell lines where KLF1 was almost completely silenced. It is not clear why cells with similar KLF1 expression levels would be discordant with regard to globin expression.

Reduction of KLF1 is expected to reverse the γ-globin to β-globin switch, and so in the cell lines where KLF1 was almost completely depleted, it was initially surprising to observe such a strong effect in the opposite direction. However, this can be explained by the fact that low levels of KLF1 are known to be required for maintenance of the chromatin architecture at the globin loci, with the activation of β-globin occurring at higher KLF1 concentrations[57,68]. Therefore the complete loss of KLF1 explains the severe reduction of highly expressed α-globin and γ-globin. This has been observed previously in K562 cells following KLF1 knockdown[57]. Unfortunately this does not account for the increased expression of β-globin, which may be occurring as a result of destabilisation of the chromatin architecture at the locus, similar to what is hypothesised in the case of the KAX9 cell line containing the ASH1L SNP.

The fact that KLF1 expression in K562 cells is minimal accounts at least in part for the phenotype of high γ-globin expression observed in these cell lines, especially since it has been shown that exogenous expression of KLF1 & BCL11A can rescue β-globin expression in K562 cells[492,493,553].

This explains the high variability observed in the rtPCR experiments targeting KLF1 in Figure 5.15, since at these low levels, even very minor fluctuations in transcript levels can result in large relative changes. The absence of KLF1 expression also provides an explanation for why a reduction in β-globin expression was not observed in the majority of cell lines containing the SNP. Since very low level KLF1 expression is required for maintenance of the chromatin architecture at the β-globin locus (1.3.1), this would also support the hypothesis that basal expression of KLF1 in K562 cells is maintaining expression of γ-globin, and that this is lost in the cell lines where KLF1 was completely ablated.

As well as being controlled at the transcriptional level, KLF1 is also subject to post-translational regulation, including SUMOylation, phosphorylation and ubiquitination[554,555]. This suggests that even the low levels of expression observed here may not accurately inform on the levels and functional activity of KLF1 protein in the cell. To accurately assess this, western blotting experiments should be conducted to inform on the affect that the KLF1 SNP has on protein levels, as well as ChIP experiments to inform on how this affects KLF1 binding activity at the globin loci.

It could also be interesting to induce differentiation of these modified K562 cell lines by treatment with hemin. Induction of differentiation in K562 cells is accompanied by increase in expression of the globin genes, as well as KLF1, and it would be interesting to see if KLF1 expression occurs to the same extent in cell lines with the SNP of interest[506].

As has been discussed, K562 cells are probably not the most useful cell system to model these SNPs *in vitro*, especially for KLF1. Now that we have successfully introduced the SNPs into K562 cells, and have improved the efficiency of the CRISPR-Cas9 pipeline in our laboratory, future work should focus on implementing this in more sensitive cell types. While this will likely require further optimisation of the technique, several potential ways to do this have been discussed in this chapter. This should probably focus on BEL-A initially, which is preferable to HUDEP-2 since it expresses β-globin without requiring differentiation, and is therefore a better model to study the impact of these SNPs on the regulation of adult haemoglobin[490]. The possibility of generating iPSCs from the patients in which the SNPs were initially identified should also be investigated, since it would be very informative to see if correcting the SNPs in these patients is sufficient to rescue β-globin expression.

6.4 Concluding Remarks

Although we were unable to conduct our investigation into the effect that HU treatment has on the epigenome of SCA patients, we propose a technique by which $CD45^{+}CD71^{+}GPA^{-}$ erythroid progenitors can be isolated directly from the peripheral blood. This technique yields sufficient DNA to perform epigenomic analyses from small quantities of blood, and will allow future longitudinal studies investigating the mechanism of action of drugs such as HU in SCA patients *in vivo*.

Through WES analyses, using the variant filtering pipeline that we developed as well as the statistical comparison of the SWiTCH and HUSTLE exome groups, we identified nine potential modifiers of SCA that warrant further investigation.

Functional validation of genomic variants *in vitro* has been facilitated by recent advances in genomic editing technologies, and provides a powerful tool to test more specifically whether a given variant is likely to affect gene function. We demonstrated that we are able to use a CRISPR-Cas9 system to successfully introduce specific mutations into cell lines, and will use this for validation of the nine candidate variants identified by the WES analyses.

The identification and validation of novel candidate variants affecting the phenotype severity, as well as mechanisms underlying the response to treatment, is crucial for advancing our understanding the complex pathophysiology of SCA, and may provide useful diagnostic indicators of risk for severe symptoms such as stroke. Particularly as Next Generation Sequencing techniques become cheaper and more easily available for use in the clinic, custom DNA capture arrays could be designed to sequence all known modifier genes in patients shortly after birth, allowing clinicians to develop customised care plans for individuals based on their genotypes.

References

1. Kister, J., Poyart, C. & Edelstein, S. J. An expanded two-state allosteric model for interactions of human hemoglobin a with nonsaturating concentrations of 2,3-diphosphoglycerate. *J. Biol. Chem.* **262,** 12085–12091 (1987).
2. Perutz, M. F. Stereochemistry of Cooperative Effects in Haemoglobin. *Nature* **228,** 726–734 (1970).
3. Sasaki, R., Ikura, K., Narita, H., Yanagawa, S. & Chiba, H. 2,3-Bisphosphoglycerate in erythroid cells. *Trends Biochem. Sci.* **7,** 140–142 (1982).
4. Yuan, Y., Tam, M. F., Simplaceanu, V. & Ho, C. New look at hemoglobin allostery. *Chem. Rev.* **115,** 1702–1724 (2015).
5. Perutz, M. F., Shih, D. T. -b. & Williamson, D. The Chloride Effect in Human Haemoglobin. *J. Mol. Biol.* **239,** 555–560 (1994).
6. Huggett, A. S. G. Foetal Blood-Gas Tensions and Gas Transfusion Through the Placenta of the Goat. *J. Physiol.* **62,** 373–384 (1927).
7. McCarthy, E. F. The oxygen affinity of human maternal and foetal haemoglobin. *J. Physiol.* **102,** 55–61 (1943).
8. Abrahamov, A. & Smith, C. A. Oxygen Capacity and Affinity of Blood from Erythroblastotic Newborns. *AMA J. Dis. Child.* **97,** 15–19 (1959).
9. Kiefer, C. M., Hou, C., Little, J. A. & Dean, A. Epigenetics of β-globin gene regulation. *Mutat. Res.* **647,** 68–76 (2008).
10. Gale, R., Clegg, J. & Huehns, E. Human embryonic haemoglobins Gower 1 and Gower 2. *Nature* **280,** 162–164 (1979).
11. Menzel, S., Garner, C., Rooks, H., Spector, T. D. & Thein, S. L. HbA2 levels in normal adults are influenced by two distinct genetic mechanisms. *Br. J. Haematol.* **160,** 101–105 (2013).
12. Felicetti, L., Novelletto, A., Benincasa, A., Terrenato, L. & Colombo, B. The HbA/HbA2 ratio in newborns and its correlation with fetal maturity. *Br. J. Haematol.* **56,** 465–71 (1984).
13. Berg, J. M., Tymoczko, J. L. & Stryer, L. in *Biochemistry* Section 10.2 (W H Freeman, 2002).
14. Olson, J. S. *et al.* The role of the distal histidine in myoglobin and haemoglobin. *Nature*

336, 265–266 (1988).

15. Friedman, J. M., Scott, T. W., Stepnoski, R. A., Ikeda-Saito, M. & Yonetani, T. The iron-proximal histidine linkage and protein control of oxygen binding in hemoglobin. A transient Raman study. *J. Biol. Chem.* **258,** 10564–10572 (1983).
16. Liddington, R., Derewenda, Z., Dodson, G. & Harris, D. Structure of the liganded T state of haemoglobin identifies the origin of cooperative oxygen binding. *Nature* **331,** 725–728 (1988).
17. Riccio, A., Vitagliano, L., di Prisco, G., Zagari, A. & Mazzarella, L. The crystal structure of a tetrameric hemoglobin in a partial hemichrome state. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 9801–6 (2002).
18. Schechter, A. N. Hemoglobin research and the origins of molecular medicine. *Blood* **112,** 3927–3938 (2008).
19. Lesecq, S. *et al.* Functional Studies and Polymerization of Recombinant Hemoglobin Glu-alpha 2beta 26(A3) → Val/Glu-7(A4) → Ala. *J. Biol. Chem.* **271,** 17211–17214 (1996).
20. Juszczak, L. J. *et al.* Conformational changes in hemoglobin S (betaE6V) imposed by mutation of the beta Glu7-beta Lys132 salt bridge and detected by UV resonance Raman spectroscopy. *J. Biol. Chem.* **278,** 7257–63 (2003).
21. Wagener, F. A. D. T. G., Abraham, N. G., van Kooyk, Y., de Witte, T. & Figdor, C. G. Heme-induced cell adhesion in the pathogenesis of sickle-cell disease and inflammation. *Trends Pharmacol. Sci.* **22,** 52–4 (2001).
22. Ballas, S. K. & Marcolina, M. J. Hyperhemolysis during the evolution of uncomplicated acute painful episodes in patients with sickle cell anemia. *Transfusion* **46,** 105–10 (2006).
23. Bunn, H. F. *et al.* Pulmonary hypertension and nitric oxide depletion in sickle cell disease. *Hypertension* **116,** 687–692 (2010).
24. Akinsheye, I. & Klings, E. S. Sickle cell anemia and vascular dysfunction: The nitric oxide connection. *J. Cell. Physiol.* **224,** 620–625 (2010).
25. Rees, D. C., Williams, T. N. & Gladwin, M. T. Sickle-cell disease. *Lancet* **376,** 2018–31 (2010).
26. Modiano, D. *et al.* Haemoglobin C protects against clinical Plasmodium falciparum malaria. *Nature* **414,** 305–308 (2001).

27. Fairhurst, R. M., Fujioka, H., Hayton, K., Collins, K. F. & Wellems, T. E. Aberrant development of Plasmodium falciparum in hemoglobin CC red cells: Implications for the malaria protective effect of the homozygous state. *Blood* **101,** 3309–3315 (2003).

28. Piel, F. B. *et al.* The distribution of haemoglobin C and its prevalence in newborns in Africa. *Sci. Rep.* **3,** (2013).

29. Cyrklaff, M. *et al.* Hemoglobins S and C Interfere with Actin Remodeling in Plasmodium falciparum–Infected Erythrocytes. *Science (80-.).* **334,** 1283–1286 (2011).

30. Charache, S., Conley, C. L., Waugh, D. F., Ugoretz, R. J. & Spurrell, J. R. Pathogenesis of hemolytic anemia in homozygous hemoglobin C disease. *J. Clin. Invest.* **46,** 1795–1811 (1967).

31. Lionnet, F. *et al.* Hemoglobin sickle cell disease complications: A clinical study of 179 cases. *Haematologica* **97,** 1136–1141 (2012).

32. Singer, K. Herditary Hemolytic Disorders Associated with Abnormal Hemoglobins. *Am. J. Med.* **18,** 633–652 (1955).

33. Hannemann, A. *et al.* The properties of red blood cells from patients heterozygous for HbS and HbC (HbSC Genotype). *Anemia* **2011,** (2011).

34. Chang, J. C. & Kan, Y. W. Beta0 Thalassemia, a Nonsense Mutation in Man. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 2886–2889 (1979).

35. Lacan, P., Aubry, M., Couprie, N. & Francina, A. Two New β^{0} -Thalassemic Mutations: A Deletion (−CC) at Codon 142 or Overlapping Codons 142-143, and an Insertion (+T) at Codon 45 or Overlapping Codons 44-45/45-46 of the β-Globin Gene. *Hemoglobin* **31,** 159–165 (2007).

36. Laosombat, V., Wongchanchailert, M., Sattayasevana, B., Wiriyasateinkul, A. & Fucharoen, S. Clinical and hematologic features of beta0-thalassemia (frameshift 41/42 mutation) in Thai patients. *Haematologica* **86,** 138–141 (2001).

37. Waye, J. S., Eng, B., Olivieri, N. F. & Chui, D. H. K. Identification of a Novel beta0-Thalassaemia Mutation in a Greek Family and Subsequent Prenatal Diagnosis. *Prenat. Diagn.* **14,** 929–932 (1994).

38. Sgourou, A. *et al.* The b-globin C to G mutation at 6 bp 3' to the termination codon causes b-thalassaemia by decreasing the mRNA level. *Br. J. Haematol.* **118,** 671–676 (2002).

39. Sgourou, A. *et al.* Thalassaemia mutations within the 5'UTR of the human beta-globin

gene disrupt transcription. *Br. J. Haematol.* **124,** 828–835 (2004).

40. Garewal, G., Das, R., Ahluwalia, J., Marwaha, R. K. & Varma, S. Nucleotide -88 (C-T) promoter mutation is a common beta-thalassemia mutation in the Jat Sikhs of Punjab, India. *Am. J. Hematol.* **79,** 252–256 (2005).

41. Cao, A. & Galanello, R. Beta-thalassemia. *Genet. Med.* **12,** 61–76 (2010).

42. Bradai, M. *et al.* Hydroxyurea can eliminate transfusion requirements in children with severe β-thalassemia. *Blood* **102,** 1529–1530 (2003).

43. Zamani, F., Shakeri, R., Eslami, S. M., Razavi, S. M. & Basi, A. Hydroxyurea therapy in 49 patients with major beta-thalassemia. *Arch. Iran. Med.* **12,** 295–297 (2009).

44. Pourfarzad, F. *et al.* Hydroxyurea responsiveness in β-thalassemic patients is determined by the stress response adaptation of erythroid progenitors and their differentiation propensity. *Haematologica* **98,** 696–704 (2013).

45. Hickman, M. *et al.* Mapping the prevalence of sickle cell and beta thalassaemia in England: estimating and validating ethnic-specific rates. *Br. J. Haematol.* **104,** 860–7 (1999).

46. Rigano, P. *et al.* Clinical and Hematological Responses to Hydroxyurea in Sicilian Patients with Hb S/β-Thalassemia. *Hemoglobin* **25,** 9–17 (2001).

47. Voskaridou, E., Kalotychou, V. & Loukopoulos, D. Clinical and laboratory effects of long-term administration of hydroxyurea to patients with sickle-cell/beta-thalassaemia. *Br. J. Haematol.* **89,** 479–484 (1995).

48. Gonzalez-Redondo, J. M. *et al.* Molecular Characterization of Hb S(C) Beta-Thalassemia in American Blacks. *Am. J. Hematol.* **38,** 9–14 (1991).

49. Lacan, P., Ponceau, B., Aubry, M. & Francina, A. Mild Hb S-beta+-Thalassemia with a Deletion of Five Nucleotides at the Polyadenylation Site of the beta-Globin Gene. *Hemoglobin* **27,** 257–259 (2003).

50. Schmugge, M., Waye, J. S., Basran, R. K., Zurbriggen, K. & Frischknecht, H. The Hb S/β+-Thalassemia Phenotype Demonstrates that the IVS-I (−2) (A>C) Mutation is a Mild β-Thalassemia Allele. *Hemoglobin* **32,** 303–307 (2008).

51. Divoky, V., Baysal, E., Schiliro, G., Dibenedetto, S. P. & Huisman, T. H. J. A Mild Type of Hb S-b+-Thalassemia [-92(C-T)] in a Sicilian family. *Am. J. Hematol.* **42,** 225–226 (1993).

52. Bresnick, E. H., Lee, H.-Y., Fujiwara, T., Johnson, K. D. & Keles, S. GATA switches as

developmental drivers. *J. Biol. Chem.* **285,** 31087–93 (2010).

53. Ting, C.-N., Olson, M. C., Barton, K. P. & Leiden, J. M. Transcription factor GATA-3 is required for development of the T-cell lineage. *Nature* **384,** 474–478 (1996).

54. Kim, Y. W., Kim, S., Kim, C. G. & Kim, A. The distinctive roles of erythroid specific activator GATA-1 and NF-E2 in transcription of the human fetal γ-globin genes. *Nucleic Acids Res.* **39,** 6944–6955 (2011).

55. Zhou, Y. *et al.* Chromatin looping defines expression of TAL1, its flanking genes, and regulation in T-ALL. *Blood* **122,** 4199–4210 (2015).

56. Kang, Y., Kim, Y. W., Kang, J., Yun, W. J. & Kim, A. Erythroid specific activator GATA-1-dependent interactions between CTCF sites around the β-globin locus. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1860,** 416–426 (2017).

57. Kang, Y., Kim, Y. W., Yun, J., Shin, J. & Kim, A. KLF1 stabilizes GATA-1 and TAL1 occupancy in the human β-globin locus. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1849,** 282–289 (2015).

58. Porcher, C., Chagraoui, H. & Kristiansen, M. S. SCL / TAL1 : a multifaceted regulator from blood development to disease. *Blood* **129,** 2051–2061 (2017).

59. Gering, M., Rodaway, A. R. F., Göttgens, B., Patient, R. K. & Green, A. R. The SCL gene specifies haemangioblast development from early mesoderm. *EMBO J.* **17,** 4029–4045 (1998).

60. Lacombe, J. *et al.* Genetic interaction between Kit and Scl. *Blood* **122,** 1150–1162 (2013).

61. Yun, W. J. *et al.* The hematopoietic regulator TAL1 is required for chromatin looping between the β-globin LCR and human γ-globin genes to activate transcription. *Nucleic Acids Res.* **42,** 4283–93 (2014).

62. Mouthon, M.-A. *et al.* Expression of tal-1 and GATA-Binding Proteins During Human Hematpoiesis. *Blood* **81,** 647–655 (1993).

63. Borg, J. *et al.* Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nat. Genet.* **42,** 801–5 (2010).

64. Zhou, D., Liu, K., Sun, C.-W., Pawlik, K. M. & Townes, T. M. KLF1 regulates BCL11A expression and gamma- to beta-globin gene switching. *Nat. Genet.* **42,** 742–4 (2010).

65. Perkins, A. C., Gaensler, K. M. & Orkin, S. H. Silencing of human fetal globin expression is impaired in the absence of the adult beta-globin gene activator protein EKLF. *Proc.*

Natl. Acad. Sci. U. S. A. **93,** 12267–71 (1996).

66. Nuez, B., Michalovich, D., Bygrave, A., Ploemacher, R. & Grosveld, F. Defective Haematopoiesis in fetal liver resulting from inactivation of the EKLF gene. *Nature* **375,** 316–318 (1995).

67. Donze, D., Townes, T. M. & Bieker, J. J. Role of Erythroid Kruppel-like Factor in Human gamma to beta Globin Gene Switching. *J. Biol. Chem.* **270,** 1955–1959 (1995).

68. Vinjamur, D. S. *et al.* Kruppel-Like transcription factor KLF1 Is required for optimal gamma- and beta-globin expression in human fetal erythroblasts. *PLoS One* **11,** 1–12 (2016).

69. Sankaran, V. G. *et al.* Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science* **322,** 1839–42 (2008).

70. Xu, J. *et al.* Transcriptional silencing of gamma-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes Dev.* **24,** 783–789 (2010).

71. Miccio, A. & Blobel, G. a. Role of the GATA-1/FOG-1/NuRD pathway in the expression of human beta-like globin genes. *Mol. Cell. Biol.* **30,** 3460–70 (2010).

72. Miccio, A. *et al.* NuRD mediates activating and repressive functions of GATA-1 and FOG-1 during blood development. *EMBO J.* **29,** 442–56 (2010).

73. Lettre, G. *et al.* DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 11869–74 (2008).

74. Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet.* **39,** 1197–9 (2007).

75. Wonkam, A. *et al.* Association of Variants at BCL11A and HBS1L-MYB with Hemoglobin F and Hospitalization Rates among Sickle Cell Patients in Cameroon. *PLoS One* **9,** e92506 (2014).

76. Cavazzana, M., Antoniani, C. & Miccio, A. Gene therapy for β-hemoglobinopathies. *Mol. Ther.* **25,** 1142–1154 (2017).

77. Bianchi, E. *et al.* c-myb supports erythropoiesis through the transactivation of KLF1 and LMO2 expression. *Blood* **116,** 99–111 (2010).

78. Gewirtz, A. M. & Calabretta, B. A c-myb antisense oligodeoxynucleotide inhibits normal human hematopoiesis in vitro. *Science (80-.).* **242,** 1303–6 (1988).

79. Stadhouders, R. *et al.* Dynamic long-range chromatin interactions control Myb proto-

oncogene transcription during erythroid development. *EMBO J.* **31,** 986–99 (2012).

80. Wahlberg, K. *et al.* The HBS1L-MYB intergenic interval associated with elevated HbF levels shows characteristics of a distal regulatory region in erythroid cells. *Blood* **114,** 1254–62 (2009).

81. Pule, G. D., Mowla, S., Novitzky, N. & Wonkam, A. Hydroxyurea down-regulates BCL11A, KLF-1 and MYB through miRNA-mediated actions to induce γ-globin expression: implications for new therapeutic approaches of sickle cell disease. *Clin. Transl. Med.* **5,** (2016).

82. Norton, L. J. *et al.* KLF1 directly activates expression of the novel fetal globin repressor ZBTB7A / LRF in erythroid cells. *Blood Adv.* **1,** 685–692 (2017).

83. Maeda, T. *et al.* LRF Is an Essential Downstream Target of GATA1 in Erythroid Development and Regulates BIM-Dependent Apoptosis. *Dev. Cell* **17,** 527–540 (2009).

84. Lunardi, A., Guarnerio, J., Wang, G., Maeda, T. & Pandolfi, P. P. Review Article Role of LRF / Pokemon in lineage fate decisions. *Blood* **121,** 2845–2853 (2013).

85. Masuda, T. *et al.* Transcription factors LRF and BCL11A independently repress expression of fetal haemoglobin. *Science (80-.).* **351,** 285–289 (2016).

86. Tanimoto, K. *et al.* Human β-Globin Locus Control Region HS5 Contains CTCF- and Developmental Activity in Erythroid Cells. *Mol. Cell. Biol.* **23,** 8946–8952 (2003).

87. Farrell, C. M., West, A. G. & Felsenfeld, G. Conserved CTCF Insulator Elements Flank the Mouse and Human beta-Globin Loci. *Mol. Cell. Biol.* **22,** 3820–3831 (2002).

88. Reik, A. *et al.* The locus control region is necessary for gene expression in the human beta-globin locus but not the maintenance of an open chromatin structure in erythroid cells [In Process Citation]. *Mol Cell Biol* **18,** 5992–6000 (1998).

89. Harju, S., Navas, P. A., Stamatoyannopoulos, G. & Peterson, K. R. Genome Architecture of the Human β-Globin Locus Affects Developmental Regulation of Gene Expression. *Mol. Cell. Biol.* **25,** 8765–8778 (2005).

90. Kim, Y. W. & Kim, A. Histone acetylation contributes to chromatin looping between the locus control region and globin gene by influencing hypersensitive site formation. *Biochim. Biophys. Acta* **1829,** 963–9 (2013).

91. Love, P. E., Warzecha, C. & Li, L. Ldb1 complexes: the new master regulators of erythroid gene transcription. *Trends Genet.* **30,** 1–9 (2014).

92. Kim, Y. W., Yun, W. J. & Kim, A. R. Erythroid activator NF-E2, TAL1 and KLF1 play roles

in forming the LCR HSs in the human adult beta-globin locus. *Int. J. Biochem. Cell Biol.* **75,** 45–52 (2016).

93. Inoue, A. *et al.* Elucidation of the role of LMO2 in human erythroid cells. *Exp. Hematol.* **41,** 1062–76.e1 (2013).

94. Wadman, I. A. *et al.* The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J.* **16,** 3145–57 (1997).

95. Krivega, I., Dale, R. K. & Dean, A. Role of LDB1 in the transition from chromatin looping to transcription activation. *Genes Dev.* **28,** 1278–90 (2014).

96. Soler, E. *et al.* The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev.* **24,** 277–89 (2010).

97. Deng, W. *et al.* Reactivation of Developmentally Silenced Globin Genes by Forced Chromatin Looping. *Cell* **158,** 849–860 (2014).

98. Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149,** 1233–1244 (2012).

99. Li, L. *et al.* Ldb1-nucleated transcription complexes function as primary mediators of global erythroid gene activation. *Blood* **121,** 4575–4585 (2013).

100. van der Ploeg, L. H. T. & Flavell, R. A. DNA methylation in the human γδβ-globin locus in erythroid and nonerythroid tissues. *Cell* **19,** 947–958 (1980).

101. Goren, A. *et al.* Fine tuning of globin gene expression by DNA methylation. *PLoS One* **1,** e46 (2006).

102. Maeder, M. L. *et al.* Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat. Biotechnol.* **31,** 1137–42 (2013).

103. Xu, J. *et al.* Corepressor-dependent silencing of fetal hemoglobin expression by BCL11A. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 6518–23 (2013).

104. Roosjen, M. *et al.* Transcriptional regulators Myb and BCL11A interplay with DNA methyltransferase 1 in developmental silencing of embryonic and fetal β-like globin genes. *FASEB J.* **28,** 1610–1620 (2014).

105. Eberharter, A. & Becker, P. B. Histone acetylation: a switch between repressive and permissive chromatin. *EMBO Rep.* **3,** 224–9 (2002).

106. Kim, Y. W. & Kim, A. Characterization of histone H3K27 modifications in the beta-globin locus. *Biochem. Biophys. Res. Commun.* **405,** 210–215 (2011).

107. Letting, D. L., Rakowski, C., Weiss, M. J. & Blobel, G. A. Formation of a Tissue-Specific Histone Acetylation Pattern by the Hematopoietic Transcription Factor GATA-1. *Mol. Cell. Biol.* **23,** 1334–1340 (2003).

108. Demers, C. *et al.* Activator-mediated recruitment of the MLL2 methyltransferase complex to the beta-globin locus. *Mol. Cell* **27,** 573–84 (2007).

109. Dzierzak, E. & Philipsen, S. Erythropoiesis: development and differentiation. *Cold Spring Harb. Perspect. Med.* **3,** a011601 (2013).

110. Sahin, A. O. & Buitenhuis, M. Molecular Mechanisms Underlying Adhesion and Migration of Hematopoietic Stem Cells. *Cell Adhes. Migr.* **6,** 39–48 (2012).

111. Golub, R. & Cumano, A. Embryonic hematopoiesis. *Blood Cells, Mol. Dis.* **51,** 226–231 (2013).

112. Oostendorp, R. A. J. & Dormer, P. VLA-4-Mediated Interactions Between Normal Human Hematopoietic Progenitors and Stromal Cells. *Leuk. Lymphoma* **24,** 423–435 (1997).

113. Lo Celso, C. & Scadden, D. T. The haematopoietic stem cell niche at a glance. *J. Cell Sci.* **124,** 3529–3535 (2011).

114. Wilson, A., Laurenti, E. & Trumpp, A. Balancing dormant and self-renewing hematopoietic stem cells. *Curr. Opin. Genet. Dev.* **19,** 461–468 (2009).

115. Greenbaum, A. *et al.* CXCL12 in early mesenchymal progenitors is required for haematopoietic stem-cell maintenance. *Nature* **495,** 227–30 (2013).

116. Nie, Y., Han, Y.-C. & Zou, Y.-R. CXCR4 is required for the quiescence of primitive hematopoietic cells. *J. Exp. Med.* **205,** 777–83 (2008).

117. Brasel, B. K. *et al.* Flt3 Ligand Synergizes With Granulocyte-Macrophage Colony-Stimulating Factor or Granulocyte Colony-Stimulating Factor to Mobilize Hematopoietic Progenitor Cells Into the Peripheral Blood of Mice. *Blood* **90,** 3781–3788 (1997).

118. Trumpp, A., Essers, M. & Wilson, A. Awakening dormant haematopoietic stem cells. *Nat. Rev. Immunol.* **10,** 201–209 (2010).

119. Kosan, C. & Godmann, M. Genetic and epigenetic mechanisms that maintain hematopoietic stem cell function. *Stem Cells Int.* **2016,** 5178965 (2016).

120. Morrison, S. J. & Weissman, I. L. The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype. *Immunity* **1,** 661–673 (1994).

121. Morrison, S. J., Wandycz, a M., Hemmati, H. D., Wright, D. E. & Weissman, I. L. Identification of a lineage of multipotent hematopoietic progenitors. *Development* **124,**

1929–1939 (1997).

122. Kondo, M. Lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors. *Immunol. Rev.* **238,** 37–46 (2010).

123. Lee, S. H. *et al.* Isolation and immunocytochemical characterization of human bone marrow stromal macrophages in hemopoietic clusters. *J. Exp. Med.* **168,** 1193–8 (1988).

124. Bessis, M. & Breton-Gorius, J. Iron Metabolism in the Bone Marrow as Seen by Electron Microscopy: A Critical Review. *1Blood* **19,** 635–663 (1962).

125. Hu, J. *et al.* Isolation and functional characterization of human erythroblasts at distinct stages: implications for understanding of normal and disordered erythropoiesis in vivo. *Blood* **121,** 3246–3253 (2013).

126. Policard, A. & Bessis, M. Micropinocytosis and Rhopheocytosis. *Nature* **194,** 110–111 (1962).

127. Brooks, R. C., Hasley, P. B., Jasti, H. & Macpherson, D. Update in general internal medicine: Evidence published in 2011. *Ann. Intern. Med.* **156,** 649–654 (2012).

128. Hanspal, M. & Hanspal, J. S. The association of erythroblasts with macrophages promotes erythroid proliferation and maturation: a 30-kD heparin-binding protein is involved in this contact. *Blood* **84,** 3494–504 (1994).

129. Wojda, U., Noel, P. & Miller, J. Fetal and adult hemoglobin production during adult erythropoiesis: coordinate expression correlates with cell proliferation. *Blood* **99,** 3005–3013 (2002).

130. Li, B., Ding, L., Li, W., Story, M. D. & Pace, B. S. Characterization of the transcriptome profiles related to globin gene switching during in vitro erythroid maturation. *BMC Genomics* **13,** 153 (2012).

131. An, X. *et al.* Global transcriptome analyses of human and murine terminal erythroid differentiation. *Blood* **123,** 3466–3477 (2014).

132. Ney, P. A. Normal and disordered reticulocyte maturation. *Curr. Opin. Hematol.* **18,** 152–157 (2011).

133. Migliaccio, A. R. Erythroblast enucleation. *Haematologica* **95,** 1985–1988 (2010).

134. Yoshida, H. *et al.* Phosphatidylserine-dependent engulfment by macrophages of nuclei from erythroid precursor cells. *Nature* **437,** 754–758 (2005).

135. Toda, S., Segawa, K. & Nagata, S. MerTK-mediated engulfment of pyrenocytes by central macrophages in erythroblastic islands. *Blood* **123,** 3963–3971 (2014).

136. Zhang, Z. W. *et al.* Red blood cell extrudes nucleus and mitochondria against oxidative stress. *IUBMB Life* **63,** 560–565 (2011).

137. Snyder, G. K. & Sheafor, B. A. Red Blood Cells : Centerpiece in the Evolution of the Vertebrate. *Amer. Zool.* **39,** 189–198 (1999).

138. Ney, P. A. Normal and disordered reticulocyte maturation. *Curr. Opin. Hematol.* **18,** 152–157 (2011).

139. Stamatoyannopoulos, G., Veith, R., Galanello, R. & Papayannopoulou, T. H. Hb F production in stressed erythropoiesis: observations and kinetic models. *Ann. N. Y. Acad. Sci.* **445,** 188–97 (1985).

140. Miller, B. A., Platt, O., Hope, S., Dover, G. & Nathan, D. G. Influence of Hydroxyurea on Fetal Hemoglobin Production in vitro. *Blood* **70,** 1824–1829 (1987).

141. Blau, C. A. *et al.* Fetal Hemoglobin in Acute and Chronic States of Erythroid Expansion. *Blood* **81,** 227–233 (1993).

142. Bard, H., Fouron, J. C., Gagnon, C. & Gagnon, J. Hypoxemia and increased fetal hemoglobin synthesis. *J. Pediatr.* **124,** 941–943 (1994).

143. Stamatoyannopoulos, G. Control of globin gene expression during development and erythroid differentiation. *Exp Hematol* **33,** 259–271 (2005).

144. Charache, S. *et al.* Hydroxyurea: effects on hemoglobin F production in patients with sickle cell anemia. *Blood* **79,** 2555–2565 (1992).

145. Xu, M. *et al.* An acetate switch regulates stress erythropoiesis. *Nat. Med.* **20,** 1018–26 (2014).

146. Menon, M. P. *et al.* Signals for stress erythropoiesis are integrated via an erythropoietin receptor-phosphotyrosine-343-Stat5 axis. *J. Clin. Invest.* **116,** 683–694 (2006).

147. Saleh, M. I., Widness, J. A. & Veng-Pederson, P. Pharmacodynamic Analysis of Stress Erythropoiesis: Change in Erythropoietin Receptor Pool Size following Double Phlebotomies in Sheep Mohammad. *Biopharm. Drug Dispos.* **32,** 131–139 (2011).

148. Jegalian, A. G., Acurio, A., Dranoff, G. & Wu, H. Erythropoietin receptor haploinsufficiency and in vivo interplay with granulocyte-macrophage colony-stimulating factor and interleukin 3. *Blood* **99,** 2603–2605 (2002).

149. Kim, T. S., Hanak, M., Trampont, P. C. & Braciale, T. J. Stress-associated erythropoiesis initiation is regulated by type 1 conventional dendritic cells. *J. Clin. Invest.* **125,** 3965–3980 (2015).

150. Luck, L., Zeng, L., Hiti, A. L., Weinberg, K. I. & Malik, P. Human CD34+ and CD34+CD38- hematopoietic progenitors in sickle cell disease differ phenotypically and functionally from normal and suggest distinct subpopulations that generate F cells. *Exp. Hematol.* **32,** 483–493 (2004).

151. Fibach, E., Manor, D., Oppenheim, A. & Rachmilewitz, E. A. Proliferation and maturation of human erythroid progenitors in liquid culture. *Blood* **73,** 100–103 (1989).

152. Horland, A. A., Wolman, S. R., Murphy, M. J. & Moore, M. A. S. Proliferation of Erythroid Colonies in Semi-Solid Agar. *Br. J. Haematol.* **36,** 495–499 (1977).

153. Leberbauer, C. *et al.* Different steroids co-regulate long-term expansion versus terminal differentiation in primary human erythroid progenitors. *Blood* **105,** 85–94 (2005).

154. Filippone, C. *et al.* Erythroid progenitor cells expanded from peripheral blood without mobilization or preselection: molecular characteristics and functional competence. *PLoS One* **5,** e9496 (2010).

155. Emerson, S. G., Thomas, S., Ferrara, J. L. & Greenstein, J. L. Developmental regulation of erythropoiesis by hematopoietic growth factors: analysis on populations of BFU-E from bone marrow, peripheral blood, and fetal liver. *Blood* **74,** 49–55 (1989).

156. Freyssinier, J. M. *et al.* Purification, amplification and characterization of a population of human erythroid progenitors. *Br. J. Haematol.* **106,** 912–922 (1999).

157. Sato, N. *et al.* In Vitro Expansion of Human Peripheral Blood CD34+ Cells. *Blood* **82,** 3600–3609 (1993).

158. Giarratana, M.-C. *et al.* Ex vivo generation of fully mature human red blood cells from hematopoietic stem cells. *Nat. Biotechnol.* **23,** 69–74 (2005).

159. van den Akker, E., Satchwell, T. J., Pellegrin, S., Daniels, G. & Toye, A. M. The majority of the in vitro erythroid expansion potential resides in CD34(-) cells, outweighing the contribution of CD34(+) cells and significantly increasing the erythroblast yield from peripheral blood samples. *Haematologica* **95,** 1594–8 (2010).

160. Munugalavadla, V. *et al.* Repression of c-kit and its downstream substrates by GATA-1 inhibits cell proliferation during erythroid maturation. *Mol. Cell. Biol.* **25,** 6747–59 (2005).

161. von Lindern, M. *et al.* The glucocorticoid receptor cooperates with the erythropoietin receptor and c-Kit to enhance and sustain proliferation of erythroid progenitors in vitro. *Blood* **94,** 550–559 (1999).

162. Koury, M. J. & Bondurant, M. C. Erythropoietin retards DNA breakdown and prevents

programmed death in erythroid progenitor cells. *Science (80-.).* **248,** 378–381 (1990).

163. Wessely, O., Deiner, E. M., Beug, H. & Von Lindern, M. The glucocorticoid receptor is a key regulator of the decision between self-renewal and differentiation in erythroid progenitors. *EMBO J.* **16,** 267–280 (1997).

164. Bauer, A. *et al.* The glucocorticoid receptor is required for stress erythropoiesis. *Genes Dev.* **13,** 2996–3002 (1999).

165. England, S. J., McGrath, K. E., Frame, J. M. & Palis, J. Immature erythroblasts with extensive ex vivo self-renewal capacity emerge from the early mammalian fetus. *Blood* **117,** 2708–2717 (2011).

166. Sawada, K., Krantz, S. B., Dessypris, E. N., Koury, S. T. & Sawyer, S. T. Human colony-forming units-erythroid do not require accessory cells, but do require direct interaction with insulin-like growth factor I and/or insulin for erythroid development. *J. Clin. Invest.* **83,** 1701–1709 (1989).

167. Correa, P. N. & Axelrad, A. A. Production of erythropoietic bursts by progenitor cells from adult human peripheral blood in an improved serum-free medium: role of insulinlike growth factor 1. *Blood* **78,** 2823–2833 (1991).

168. Muta, K., Krantz, S. B., Bondurant, M. C. & Wickrema, A. Distinct roles of erythropoietin, insulin-like growth factor I, and stem cell factor in the development of erythroid progenitor cells. *J. Clin. Invest.* **94,** 34–43 (1994).

169. Migliaccio, G. *et al.* In Vitro Mass Production of Human Erythroid Cells from the Blood of Normal Donors and of Thalassemic Patients. *Blood Cells, Mol. Dis.* **28,** 169–180 (2002).

170. Chao, R., Gong, X., Wang, L., Wang, P. & Wang, Y. CD71high population represents primitive erythroblasts derived from mouse embryonic stem cells. *Stem Cell Res.* **14,** 3038 (2015).

171. Rojas-Sutterlin, S., Lecuyer, E. & Hoang, T. Kit and Scl regulation of hematopoietic stem cells. *Curr. Opin. Hematol.* **21,** 256–64 (2014).

172. Mendelson, A. & Frenette, P. S. Hematopoietic stem cell niche maintenance during homeostasis and regeneration. *Nat. Med.* **20,** 833–846 (2014).

173. Munugalavadla, V. *et al.* Repression of c-kit and its downstream substrates by GATA-1 inhibits cell proliferation during erythroid maturation. *Mol. Cell. Biol.* **25,** 6747–59 (2005).

174. Li, J. *et al.* Isolation and transcriptome analyses of human erythroid progenitors: BFU-E and CFU-E. *Blood* **124,** 3636–3645 (2014).

175. Migliaccio, G. *et al.* In Vitro Mass Production of Human Erythroid Cells from the Blood of Normal Donors and of Thalassemic Patients. *Blood Cells, Mol. Dis.* **28,** 169–180 (2002).

176. Jersmann, H. P. A. Time to abandon dogma: CD14 is expressed by non-myeloid lineage cells. *Immunol. Cell Biol.* **83,** 462–467 (2005).

177. Herrick, J. B. Peculiar Elongated and Sickle-Shaped Red Blood Corpuscles in a Case of Severe Anaemia. *Arch. Intern. Med.* **VI,** 517–521 (1910).

178. Washburn, R. E. Peculiar Elongated and Sickle-Shaped Red Blood Corpuscles in a Case of Severe Anaemia. *Viginia Med. Semi-Monthly* **15,** 490–493 (1911).

179. Cook, J. E. & Meyer, J. Severe Anemia with Remarkable Elongated and Sickle-Shaped Red Blood Cells and Chronic Leg Ulcer. *Arch. Intern. Med.* **XVI,** 644–651 (1915).

180. Mason, V. R. Sickle Cell Anemia. *JAMA* **79,** 1318–1320 (1922).

181. Beet, E. A. The genetics of the sickle-cell trait in a Bantu tribe. *Ann. Eugen.* **14,** 279–284 (1949).

182. Neel, J. V. The Inheritance of Sickle Cell Anemia. *Science (80-.).* **110,** 64–66 (1949).

183. Pauling, L., Itano, H. A., Singer, S. J. & Wells, I. C. Sickle Cell Anemia, a Molecular Disease. *Science (80-.).* **110,** 543–548 (1949).

184. Ingram, V. M. A specific chemical difference between the globins of normal human and sickle-cell anæmia hæmoglobin. *Nature* **178,** 792–794 (1956).

185. Kan, Y. W. & Dozy, A. M. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proc. Natl. Acad. Sci. U. S. A.* **75,** 5631–5 (1978).

186. Kan, Y. W. & Dozy, A. M. ANTENATAL DIAGNOSIS OF SICKLE-CELL ANAEMIA BY D.N.A. ANALYSIS OF AMNIOTIC-FLUID CELLS. *Lancet* **312,** 910–912 (1978).

187. Gabriel, A. & Przybylski, J. Sickle-Cell Anemia: A Look at Global Haplotype Distribution. *Nature Education* **3,** (2010).

188. Friedman, M. J. Erythrocytic mechanism of sickle cell resistance to malaria. *Proc. Natl. Acad. Sci. U. S. A.* **75,** 1994–7 (1978).

189. Williams, T. N. *et al.* An immune basis for malaria protection by the sickle cell trait. *PLoS Med.* **2,** e128 (2005).

190. Piel, F. B. *et al.* Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat. Commun.* **1,** 104 (2010).

191. 1000 Genomes Project Consortium. A global reference for human genetic variation.

Nature **526,** 68–74 (2015).

192. Mcauley, C. F. *et al.* High mortality from Plasmodium falciparum malaria in children living with sickle cell anemia on the coast of Kenya. **116,** 1663–1669 (2015).

193. Luzzatto, L. Sickle cell anaemia and malaria. *Mediterr. J. Hematol. Infect. Dis.* **4,** e2012065 (2012).

194. Luzzatto, L., Nwachuku-Jarrett, E. S. & Reddy, S. INCREASED SICKLING OF PARASITISED ERYTHROCYTES AS MECHANISM OF RESISTANCE AGAINST MALARIA IN THE SICKLE-CELL TRAIT. *Lancet* **295,** 319–322 (1970).

195. Ayi, K., Turrini, F., Piga, A. & Arese, P. Enhanced phagocytosis of ring-parasitized mutant erythrocytes: A common mechanism that may explain protection against falciparum malaria in sickle trait and beta-thalassemia trait. *Blood* **104,** 3364–3371 (2004).

196. Mohandas, N. & An, X. Malaria and Human Red Blood Cells. *Med Microbiol Immunol.* **201,** 593–598 (2012).

197. NHS Choices: Sickle cell anaemia. (2010). Available at: http://www.nhs.uk/conditions/Sickle-cell-anaemia/Pages/Introduction.aspx.

198. Pizzo, E. *et al.* A retrospective analysis of the cost of hospitalizations for sickle cell disease with crisis in England, 2010/11. *J. Public Health (Oxf).* 1–11 (2014). doi:10.1093/pubmed/fdu026

199. Piel, F. B. *et al.* Global migration and the changing distribution of sickle haemoglobin: a quantitative study of temporal trends between 1960 and 2000. *lancet Glob. Heal.* **2,** e80–e89 (2014).

200. Powars, D. R., Chan, L. S., Hiti, A., Ramicone, E. & Johnson, C. Outcome of Sickle Cell Anemia: A 4-Decade Observational Study of 1056 Patients. *Medicine (Baltimore).* **84,** 363–376 (2005).

201. Schimmel, M. *et al.* Nucleosomes and neutrophil activation in sickle cell disease painful crisis. *Haematologica* **98,** 1797–803 (2013).

202. Solovey, A. *et al.* Circulating activated endothelial cells in sickle cell anemia. *N. Engl. J. Med.* **337,** 1584–90 (1997).

203. Kaul, D. K. & Hebbel, R. P. Hypoxia/reoxygenation causes inflammatory response in transgenic sickle mice but not in normal mice. *J. Clin. Invest.* **106,** 411–20 (2000).

204. Babadoko, A. A. *et al.* Autosplenectomy of sickle cell disease in zaria, Nigeria: an

ultrasonographic assessment. *Oman Med. J.* **27,** 121–3 (2012).

205. Nottage, K. A. *et al.* Predictors Of Splenic Function Preservation In Children With Sickle Cell Anaemia Treated With Hydroxyurea. *Eur. J. Haematol.* **93,** 377–383 (2014).

206. Booth, C., Inusa, B. & Obaro, S. K. Infection in sickle cell disease: A review. *Int. J. Infect. Dis.* **14,** 2–12 (2010).

207. Arkuszewski, M. *et al.* Sickle cell anemia: Intracranial stenosis and silent cerebral infarcts in children with low risk of stroke. *Adv. Med. Sci.* **59,** 108–13 (2014).

208. Steinberg, M. Predicting clinical severity in sickle cell anaemia. *Br. J. Haematol.* **129,** 465–81 (2005).

209. Marziali, M. *et al.* Peripheral red blood cell split chimerism as a consequence of intramedullary selective apoptosis of recipient red blood cells in a case of sickle cell disease. *Mediterr. J. Hematol. Infect. Dis.* **6,** e2014066 (2014).

210. Walters, M. C. *et al.* Impact of bone marrow transplantation for symptomatic sickle cell disease: an interim report. *Blood* **95,** 1918–1924 (2000).

211. Panepinto, J. A. *et al.* Matched-related donor transplantation for sickle cell disease: Report from the Center for International Blood and Transplant Research. *Br. J. Haematol.* **137,** 479–485 (2007).

212. Mentzer, W., Heller, S., Pearle, P., Hackney, E. & Vichinsky, E. Availability of Related Donors in Sickle Cell Anemia. *Am. J. Pediatr. Hematol. Oncol.* **16,** 27–29 (1994).

213. Bolaños-Meade, J. & Brodsky, R. A. Blood and marrow transplantation for sickle cell disease: overcoming barriers to success. *Curr. Opin. Oncol.* **21,** 158–61 (2009).

214. Kassim, A. A. & Debaun, M. R. The case for and against initiating either hydroxyurea therapy, blood transfusion therapy or hematopoietic stem cell transplant in asymptomatic children with sickle cell disease. *Expert Opin. Pharmacother.* **15,** 325–336 (2014).

215. Adams, R. J. *et al.* Prevention of a First Stroke by Transfusions in Children with Sickle Cell Anemia and Abnormal Results on Transcranial Doppler Ultrasonography. *N. Engl. J. Med.* **339,** 5–11 (1998).

216. Adams, R. *et al.* The Use of Transcranial Ultrasonography to Predict Stroke in Sickle Cell Disease. *N. Engl. J. Med.* **326,** 605–610 (1992).

217. Howard, J. The role of blood transfusion in Sickle Cell Disease. *ISBT Sci. Ser.* **8,** 225–228 (2013).

218. Lee, M. T. *et al.* Stroke Prevention Trial in Sickle Cell Anemia (STOP): extended follow-

up and final results. *Blood* **108,** 847–852 (2006).

219. Pandey, H., Das, S. S. & Chaudhary, R. Red cell alloimmunization in transfused patients: A silent epidemic revisited. *Asian J. Transfus. Sci.* **8,** 75–77 (2014).

220. Olujohungbe, A., Hambleton, I., Stephens, L., Serjeant, B. & Serjeant, G. Red cell antibodies in patients with homozygous sickle cell disease: a comparison of patients in Jamaica and the United Kingdom. *Br. J. Haematol.* **113,** 661–665 (2001).

221. Singer, S. T. *et al.* Alloimmunization and erythrocyte autoimmunization in transfusion-dependent thalassemia patients of predominantly asian descent. *Blood* **96,** 3369–3373 (2000).

222. Porter, J. & Garbowski, M. Consequences and management of iron overload in sickle cell disease. *Hematology* **2013,** 447–56 (2013).

223. Meloni, A. *et al.* Cardiac iron overload in sickle-cell disease. *Am. J. Hematol.* **89,** 678–683 (2014).

224. Darbari, D. S. *et al.* Circumstances of Death in Adult Sickle Cell Disease Patients. *Am. J. Hematol.* **81,** 858–863 (2006).

225. Aduloju, S. O., Palmer, S. & Eckman, J. R. Mortality in Sickle Cell Patient Transitioning from Pediatric to Adult Program: 10 Years Grady Comprehensive Sickle Cell Center Experience. *Blood* **112,** 1426 (2008).

226. Charache, S. *et al.* Effect of Hydroxyurea on the Frequency of Painful Crises in Sickle Cell. *N. Engl. J. Med.* **332,** 1317–1322 (1995).

227. Wang, W. C. *et al.* Hydroxycarbamide in very young children with sickle-cell anaemia: A multicentre, randomised, controlled trial (BABY HUG). *Lancet* **377,** 1663–1672 (2011).

228. Steinberg, M. *et al.* The Risks and Benefits of Long-term Use of Hydroxyurea in Sickle Cell Anemia: A 17.5 Year Follow-Up. *Am J Hematol* **85,** 403–408 (2011).

229. Le, P. Q. *et al.* Survival Among Children and Adults with Sickle Cell Disease in Belgium: Benefit from Hydroxyurea Treatment. *Pediatr. Blood Cancer* **62,** 1956–1961 (2015).

230. Lopes de Castro Lobo, C. *et al.* The effect of hydroxcarbamide therapy on survival of children with sickle cell disease. *Br. J. Haematol.* **161,** 852–860 (2013).

231. West, W. O. Hydroxyurea in the treatment of Polycythemia Vera: A prospective study of 100 patients over a 20-year period. *South. Med. J.* **80,** 323–327 (1987).

232. Platt, O. S. *et al.* Hydroxyurea enhances fetal hemoglobin production in sickle cell anemia. *J. Clin. …* **74,** 652–656 (1984).

233. Letvin, N. L., Linch, D. C., Beardsley, G. P., McIntyre, K. W. & Nathan, D. G. Augmentation of Fetal-Hemoglobin Production in Anaemic Monkeys by Hydroxyurea. *N. Engl. J. Med.* **310,** 869–873 (1984).

234. Yarbro, J. W., Kennedy, B. J. & Barnum, C. P. Hydroxyurea Inhibition of DNA Synthesis in Ascites Tumor. *Proc. Natl. Acad. Sci. USA* **53,** 1033–1035 (1965).

235. Young, C. W. & Hodas, S. Hydroxyurea: Inhibitory Effect on DNA Metabolism. *Science (80-.).* **146,** 1172–1174 (1964).

236. Koç, A., Wheeler, L. J., Mathews, C. K. & Merrill, G. F. Hydroxyurea Arrests DNA Replication by a Mechanism that Preserves Basal dNTP Pools. *J. Biol. Chem.* **279,** 223–230 (2004).

237. Kolberg, M., Strand, K. R., Graff, P. & Andersson, K. K. Structure, function, and mechanism of ribonucleotide reductases. *Biochim. Biophys. Acta - Proteins Proteomics* **1699,** 1–34 (2004).

238. Ballas, S. K., Marcolina, M. J., Dover, G. J. & Barton, F. B. Erythropoietic activity in patients with sickle cell anaemia before and after treatment with hydroxyurea. *Br. J. Haematol.* **105,** 491–496 (1999).

239. Flanagan, J. M. *et al.* Hydroxycarbamide alters erythroid gene expression in children with sickle cell anaemia. *Br. J. Haematol.* **157,** 240–8 (2012).

240. Grieco, A. J., Billett, H. H., Green, N. S., Driscoll, M. C. & Bouhassira, E. E. Variation in gamma-globin expression before and after induction with hydroxyurea associated with BCL11A, KLF1 and TAL1. *PLoS One* **10,** e0129431 (2015).

241. Cokic, V. P. *et al.* Hydroxyurea induces fetal hemoglobin by the nitric oxide–dependent activation of soluble guanylyl cyclase. *J. Clin. Invest.* **111,** 231–239 (2003).

242. King, S. B. Nitric oxide production from hydroxyurea. *Free Radic. Biol. Med.* **37,** 737–44 (2004).

243. Huang, J., Yakubu, M., Kim-Shapiro, D. B. & King, S. B. Rat liver-mediated metabolism of hydroxyurea to nitric oxide. *Free Radic. Biol. Med.* **40,** 1675–81 (2006).

244. Lockwood, S. Y., Erkal, J. L. & Spence, D. M. Endothelium-derived nitric oxide production is increased by ATP released from red blood cells incubated with hydroxyurea. *Nitric Oxide* **30,** 1–7 (2014).

245. Walker, A. L. *et al.* Epigenetic and molecular profiles of erythroid cells after hydroxyurea treatment in sickle cell anemia. *Blood* **118,** 5664–70 (2011).

246. Ronchi, A. & Ottolenghi, S. To respond or not to respond to hydroxyurea in thalassemia: A matter of stress adaptation? *Haematologica* **98,** 657–659 (2013).

247. Mabaera, R. *et al.* A cell stress signaling model of fetal hemoglobin induction: what doesn't kill red blood cells may make them stronger. *Exp. Hematol.* **36,** 1057–1072 (2008).

248. Silva-Pinto, A. C. *et al.* Hydroxycarbamide modulates components involved in the regulation of adenosine levels in blood cells from sickle-cell anemia patients. *Ann. Hematol.* (2014). doi:10.1007/s00277-014-2066-4

249. Styles, L. A. *et al.* Decrease of Very Late Activation Antigen-4 and CD36 on Reticulocytes in Sickle Cell Patients Treated With Hydroxyurea. *Blood* **89,** 2554–2559 (1997).

250. Ware, R. E. & Helms, R. W. Stroke With Transfusions Changing to Hydroxyurea (SWiTCH). *Blood* **119,** 3925–3932 (2012).

251. Ware, R. E. *et al.* Stroke With Transfusions Changing to Hydroxyurea (SWiTCH): A Phase III Randomized Clinical Trial for Treatment of Children With Sickle Cell Anemia, Stroke, and Iron Overload. *Pediatr. Blood Cancer* **57,** 1011–1017 (2011).

252. Ware, R. E. *et al.* Hydroxycarbamide versus chronic transfusion for maintenance of transcranial doppler flow velocities in children with sickle cell anaemia - TCD with Transfusions Changing to Hydroxyurea (TWiTCH): A multicentre, open-label, phase 3, non-inferiority trial. *Lancet* **387,** 661–670 (2016).

253. Perrine, S. P. *et al.* A Short-Term Trial of Butyrate to Stimulate Fetal-Globin Gene Expression in the β-Globin Disorders. *N. Engl. J. Med.* **328,** 81–86 (1993).

254. DeSimone, J., Heller, P., Hall, L. & Zwiers, D. 5-Azacytidine stimulates fetal hemoglobin synthesis in anemic baboons. *Proc. Natl. Acad. Sci. U. S. A.* **79,** 4428–31 (1982).

255. Ley, T. J. *et al.* 5-Azacytidine Selectively Increases γ-Globin Synthesis in a Patient with β+ Thalassemia. *N. Engl. J. Med.* **307,** 1469–1475 (1982).

256. Charache, S. *et al.* Treatment of sickle cell anemia with 5-azacytidine results in increased fetal hemoglobin production and is associated with nonrandom hypomethylation of DNA around the gamma-delta-beta-globin gene complex. *Proc. Natl. Acad. Sci. U. S. A.* **80,** 4842–6 (1983).

257. Weinberg, R. S. *et al.* Butyrate increases the efficiency of translation of gamma-globin mRNA. *Blood* **105,** 1807–1809 (2005).

258. Dulmovits, B. M. *et al.* Pomalidomide reverses gamma-globin silencing through the transcriptional reprogramming of adult hematopoietic progenitors. *Blood* **127,** 1481–1492 (2016).

259. Pecoraro, A. *et al.* Efficacy of Rapamycin as Inducer of Hb F in Primary Erythroid Cultures from Sickle Cell Disease and beta-Thalassemia Patients. *Hemoglobin* **39,** 225–229 (2015).

260. Hannemann, A., Cytlak, U. M., Rees, D. C., Tewari, S. & Gibson, J. S. Effects of 5-hydroxymethyl-2-furfural on the volume and membrane permeability of red blood cells from patients with sickle cell disease. *J. Physiol.* **592,** 4039–4049 (2014).

261. Oder, E., Safo, M. K., Abdulmalik, O. & Kato, G. J. New developments in anti-sickling agents: can drugs directly prevent the polymerization of sickle haemoglobin in vivo? *Br. J. Haematol.* **175,** 24–30 (2016).

262. Morris, C. R. *et al.* A randomized, placebo-controlled trial of arginine therapy for the treatment of children with sickle cell disease hospitalized with vaso-occlusive pain episodes. *Haematologica* **98,** 1375–82 (2013).

263. Kosinski, P., Croal, P., Leung, J., Williams, S. & Kassner, A. Assessing the Effect of Short and Long-Term Hydroxyurea Treatment on Cerebral Hemodynamics in Children with Sickle Cell Anemia Using Quantitative MRI: Preliminary Findings. *Blood* **124,** 4090 (2014).

264. Diaz-Perez, F. *et al.* L-arginine transport and nitric oxide synthesis in human endothelial progenitor cells. *J. Cardiovasc. Pharmacol.* **60,** 439–449 (2012).

265. NHS. NHS Choices: Treatments for Sickle Cell Disease. *NHS Choices* (2016). Available at: http://www.nhs.uk/Conditions/Sickle-cell-anaemia/Pages/Treatment.aspx#infections.

266. Gaston, M. H. *et al.* Prophylaxis with Oral Penicillin in Children with Sickle Cell Anemia: A Randomized Trial. *N. Engl. J. Med.* **314,** 1593–1599 (1986).

267. Sobota, A., Sabharwal, V., Fonebi, G. & Steinberg, M. How we prevent and manage infection in sickle cell disease. *Br. J. Haematol.* **170,** 757–767 (2015).

268. Hirst, C. & Owusu-Ofori, S. Prophylactic antibiotics for preventing pneumococcal infection in children with sickle cell disease. *Cochrane database Syst. Rev.* CD003427 (2012). doi:10.1002/14651858.CD003427

269. Akinsheye, I. *et al.* Fetal hemoglobin in sickle cell anemia. *Blood* **118,** 19–27 (2011).

270. Elenga, N. *et al.* Pregnancy in Sickle Cell Disease Is a Very High-Risk Situation: An

Observational Study. *Obstet. Gynecol. Int.* **2016,** 1–5 (2016).

271. De Montalembert, M. & Deneux-Tharaux, C. Pregnancy in sickle cell disease is at very high risk. *Blood* **125,** 3216–3218 (2015).

272. Quinn, C. T. Sickle Cell Disease in Childhood. *Pediatr. Clin. North Am.* **60,** 1363–1381 (2013).

273. Boyer, S. H., Belding, T. K., Margolet, L. & Noyes, a N. Fetal hemoglobin restriction to a few erythrocytes (F cells) in normal human adults. *Science (80-.).* **188,** 361–363 (1975).

274. Zertal-Zidani, S., Ducrocq, R., Sahbatou, M., Satta, D. & Krishnamoorthy, R. Foetal haemoglobin in normal healthy adults: relationship with polymorphic sequences cis to the beta globin gene. *Eur. J. Hum. Genet.* **10,** 320–6 (2002).

275. Toma, S., Tenorio, M., Oakley, M., Thein, S. L. & Clark, B. E. Two Novel Mutations (*HBG1* : c.-250C>T and *HBG2* : c.-250C>T) Associated With Hereditary Persistence of Fetal Hemoglobin. *Hemoglobin* **38,** 67–69 (2014).

276. Thein, S. L., Menzel, S., Lathrop, M. & Garner, C. Control of fetal hemoglobin: new insights emerging from genomics and clinical implications. *Hum. Mol. Genet.* **18,** R216-23 (2009).

277. Hariharan, P., Gorivale, M., Colah, R., Ghosh, K. & Nadkarni, A. Does the Novel KLF1 Gene Mutation Lead to a Delay in Fetal Hemoglobin Switch? *Ann. Hum. Genet.* **81,** 125–128 (2017).

278. Bank, A. Regulation of human fetal hemoglobin: new players, new complexities. *Blood* **107,** 435–443 (2006).

279. O'Neil, J. *et al.* Alu elements mediate MYB gene tandem duplication in human T-ALL. *J. Exp. Med.* **204,** 3059–66 (2007).

280. Ramsay, R. G. & Gonda, T. J. MYB function in normal and cancer cells. *Nat. Rev. Cancer* **8,** 523–34 (2008).

281. Thein, S. L. *et al.* Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 11346–51 (2007).

282. Loggetto, S. R. Sickle cell anemia: clinical diversity and beta S-globin haplotypes. *Rev. Bras. Hematol. Hemoter.* **35,** 155–157 (2013).

283. Creary, L. E. *et al.* Genetic variation on chromosome 6 influences F cell levels in healthy individuals of African descent and HbF levels in sickle cell patients. *PLoS One* **4,** e4218

(2009).

284. Embury, S. H. *et al.* Two different molecular organizations account for the single alpha-globin gene of the alpha-thalassemia-2 genotype. *J. Clin. Invest.* **66,** 1319–25 (1980).

285. Galanello, R. & Cao, A. Alpha-thalassemia. *Genet. Med.* **13,** 83–88 (2011).

286. Higgs, D. R. & Weatherall, D. J. The Alpha Thalassaemias. *Cell. Mol. Life Sci.* **66,** 1154–1162 (2009).

287. Williams, T. N. *et al.* Both heterozygous and homozygous alpha+ thalassemias protect against severe and fatal Plasmodium falciparum malaria on the coast of Kenya. *Blood* **106,** 368–371 (2005).

288. Mockenhaupt, F. P. *et al.* alpha+-thalassemia protects African children from severe malaria. *Blood* **104,** 2003–2006 (2004).

289. Allen, S. J. *et al.* alpha+-Thalassemia protects children against disease caused by other infections as well as malaria. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 14736–41 (1997).

290. Lubega, I., Ndugwa, C. M., Mworozi, E. A. & Tumwine, J. K. Alpha thalassemia among sickle cell anaemia patients in Kampala, Uganda. *Afr. Health Sci.* **15,** 682–689 (2015).

291. Saleh-Gohari, N. & Mohammadi-Anaie, M. Co-inheritance of sickle cell trait and Thalassemia mutations in South Central Iran. *Iran. J. Public Health* **41,** 81–86 (2012).

292. Steinberg, M. H. & Embury, S. H. Alpha-thalassemia in blacks: genetic and clinical aspects and interactions with the sickle hemoglobin gene. *Blood* **68,** 985–990 (1986).

293. Wonkam, A. *et al.* Co-inheritance of sickle cell anemia and α-thalassemia delays disease onset and could improve survival in Cameroonian's patients (sub-Saharan Africa). *Am. J. Hematol.* **89,** 664–665 (2014).

294. Embury, S. H., Clark, M. R., Monroy, G. & Mohandas, N. Concurrent sickle cell anemia and a-thalassemia. *J. Clin. Invest.* **73,** 116–123 (1984).

295. Cox, S. E. *et al.* Haptoglobin, alpha-thalassaemia and glucose-6-phosphate dehydrogenase polymorphisms and risk of abnormal transcranial Doppler among patients with sickle cell anaemia in Tanzania. *Br. J. Haematol.* (2014). doi:10.1111/bjh.12791

296. Rumaney, M. B. *et al.* The co-inheritance of alpha-thalassemia and sickle cell anemia is associated with better hematological indices and lower consultations rate in cameroonian patients and could improve their survival. *PLoS One* **9,** e100516 (2014).

297. Domingos, I. F. *et al.* Influence of the β(s) haplotype and α-thalassemia on stroke

development in a Brazilian population with sickle cell anaemia. *Ann. Hematol.* 10–12 (2014). doi:10.1007/s00277-014-2016-1

298. Pandey, S. K. *et al.* Phenotypic Effect of α-Globin Gene Numbers on Indian Sickle β-Thalassemia Patients. *J. Clin. Lab. Anal.* **4,** 1–4 (2014).

299. Embury, S. H. Alpha Thalassemia - a Modifier of Sickle-Cell Disease. *Ann. N. Y. Acad. Sci.* **565,** 213–221 (1989).

300. Steinberg, M. H. Genetic Modifiers of Sickle Cell Disease. *Am. J. Hematol.* **87,** 795–803 (2012).

301. Vasavda, N. *et al.* Effects of co-existing α-thalassaemia in sickle cell disease on hydroxycarbamide therapy and circulating nucleic acids. *Br. J. Haematol.* **157,** 249–52 (2012).

302. Amin, B. R. *et al.* Monozygotic twins with sickle cell anemia and discordant clinical courses: clinical and laboratory studies. *Hemoglobin* **15,** 247–256 (1991).

303. Joishy, S. K., Griner, P. F. & Rowley, P. T. Sickle β-thalassemia: Identical twins differing in severity implicate nongenetic factors influencing course. *Am. J. Hematol.* **1,** 23–33 (1976).

304. Weatherall, M. W., Higgs, D. R., Weiss, H., Weatherall, D. J. & Serjeant, G. R. Phenotype/genotype relationships in sickle cell disease: a pilot twin study. *Clin. Lab. Haematol.* **27,** 384–90 (2005).

305. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J. Bacteriol.* **169,** 5429–33 (1987).

306. Mojica, F. J. M., Juez, G. & Rodriguez-Valera, F. Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. *Mol. Microbiol.* **9,** 613–621 (1993).

307. Jansen, R., Van Embden, J. D. A., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43,** 1565–1575 (2002).

308. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60,** 174–182 (2005).

309. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire

new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151,** 653–663 (2005).

310. Bolotin, A., Quinquis, B., Sorokin, A. & Dusko Ehrlich, S. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151,** 2551–2561 (2005).

311. Lander, E. S. The Heroes of CRISPR. *Cell* **164,** 18–28 (2016).

312. Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokarytotes. *Science (80-.).* **315,** 1709–1712 (2007).

313. Brouns, S. J. J. *et al.* Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science (80-.).* **321,** 619–621 (2008).

314. Marraffini, L. A. & Sontheimer, E. J. CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. *Science (80-.).* **322,** 1843–1845 (2008).

315. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468,** 67–71 (2010).

316. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–607 (2011).

317. Jinek, M. *et al.* Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. *Science (80-.).* **343,** 1247997–1247997 (2014).

318. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E2579-86 (2012).

319. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (80-.).* **337,** 816–821 (2012).

320. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (80-.).* **339,** 819–823 (2013).

321. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science (80-.).* **339,** 823–826 (2013).

322. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *Elife* **2013,** e00471 (2013).

323. Cho, S. W., Kim, S., Kim, J. M. & Kim, J.-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **31,** 230–2 (2013).

324. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology.

Nat. Methods **10,** 957–63 (2013).

325. Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in type II-A CRISPR-Cas adaptation. *Genes Dev.* **29,** 356–361 (2015).

326. Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519,** 199–202 (2015).

327. Klein, L., Kyewski, B., Allen, P. M. & Hogquist, K. A. Positive and negative selection of the T cell repertoire: what thymocytes see and don't see. *Nat. Rev. Immunol.* **14,** 377–391 (2014).

328. Nuñez, J. K. *et al.* Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21,** 528–34 (2014).

329. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13,** 722–736 (2015).

330. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31,** 833–8 (2013).

331. Shen, B. *et al.* Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. *Nat. Methods* **11,** 399–402 (2014).

332. Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33,** 510–517 (2015).

333. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* **10,** 1116–21 (2013).

334. Xu, X. *et al.* A CRISPR-based approach for targeted DNA demethylation. *Cell Discov.* **2,** doi:10.1038/celldisc.2016.9 (2016).

335. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163,** 759–771 (2015).

336. Lans, H., Marteijn, J. A. & Vermeulen, W. ATP-dependent chromatin remodeling in the DNA-damage response. *Epigenetics Chromatin* **5,** (2012).

337. Sartori, A. A. *et al.* Human CtIP promotes DNA end resection. *Nature* **450,** 509–514 (2007).

338. Jasin, M. & Rothstein, R. Repair of strand breaks by homologous recombination. *Cold Spring Harb. Perspect. Biol.* **5,** 1–19 (2013).

339. Soldner, F. *et al.* Generation of isogenic pluripotent stem cells differing exclusively at two early onset parkinson point mutations. *Cell* **146,** 318–331 (2011).

340. Song, F. & Stieger, K. Optimizing the DNA Donor Template for Homology-Directed Repair of Double-Strand Breaks. *Mol. Ther. - Nucleic Acids* **7,** 53–60 (2017).

341. Choulika, A., Perrin, A., Dujon, B. & Ois Nicolas, J. Induction of Homologous Recombination in Mammalian Chromosomes by Using the I-SceI System of Saccharomyces cerevisiae. *Mol. Cell. Biol.* **15,** 1968–1973 (1995).

342. Maeder, M. L. & Gersbach, C. A. Genome-editing technologies for gene and cell therapy. *Mol. Ther.* **24,** 430–446 (2016).

343. Chevalier, B. S. & Stoddard, B. L. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res.* **29,** 3757–74 (2001).

344. Porteus, M. H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science (80-.).* **300,** 763 (2003).

345. Bibikova, M., Golic, M., Golic, K. G. & Carroll, D. Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. *Genetics* **161,** 1169–1175 (2002).

346. Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29,** 143–150 (2011).

347. Chen, F. *et al.* High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat. Methods* **8,** 753–5 (2011).

348. Hirsch, M. L., Green, L., Porteus, M. H. & Samulski, R. J. Self-complementary AAV mediates gene targeting and enhances endonuclease delivery for double-strand break repair. *Gene Ther.* **17,** 1175–1180 (2010).

349. Urnov, F. D. *et al.* Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435,** 646–651 (2005).

350. Moehle, E. A. *et al.* Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. *Proc. Natl. Acad. Sci.* **104,** 3055–3060 (2007).

351. Zhang, J. P. *et al.* Efficient precise knockin with a double cut HDR donor after CRISPR/Cas9-mediated double-stranded DNA cleavage. *Genome Biol.* **18,** 1–18 (2017).

352. Bertolini, L. R., Bertolini, M., Maga, E. A., Madden, K. R. & Murray, J. D. Increased gene targeting in Ku70 and Xrcc4 transiently deficient human somatic cells. *Mol. Biotechnol.* **41,** 106–114 (2009).

353. Ho, T. T. *et al.* Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell

lines. *Nucleic Acids Res.* **43,** e17 (2015).

354. Zhu, L., Mon, H., Xu, J., Lee, J. M. & Kusakabe, T. CRISPR / Cas9-mediated knockout of factors in non-homologous end joining pathway enhances gene targeting in silkworm cells. *Nat. Publ. Gr.* 1–13 (2015). doi:10.1038/srep18103

355. Difilippantonio, M. J. *et al.* Dna repair protein Ku80 suppresses chromosomal aberrations and malignant transformation. *Nature* **404,** 510–514 (2000).

356. Ngo, J. *et al.* Bax deficiency extends the survival of Ku70 knockout mice that develop lung and heart diseases. *Cell Death Dis.* **6,** e1706 (2015).

357. Srivastava, M. *et al.* An inhibitor of nonhomologous end-joining abrogates double-strand break repair and impedes cancer progression. *Cell* **151,** 1474–1487 (2012).

358. Chu, V. T. *et al.* letters Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.* **33,** 543–548 (2015).

359. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle* **7,** 2902–2906 (2008).

360. Gutschner, T., Haemmerle, M., Genovese, G., Draetta, G. F. & Chin, L. Post-translational Regulation of Cas9 during G1 Enhances Homology-Directed Repair. *Cell Rep.* **14,** 1555–1566 (2016).

361. Howden, S. E. *et al.* A Cas9 Variant for Efficient Generation of Indel-Free Knockin or Gene-Corrected Human Pluripotent Stem Cells. *Stem Cell Reports* **7,** 508–517 (2016).

362. Schaefer, K. A. *et al.* Unexpected mutations after CRISPR-Cas9 editing in vivo. *Nat. Methods* **14,** 547–548 (2017).

363. Zhang, X. H., Tee, L. Y., Wang, X. G., Huang, Q. S. & Yang, S. H. Off-target effects in CRISPR/Cas9-mediated genome engineering. *Mol. Ther. - Nucleic Acids* **4,** e264 (2015).

364. Cameron, P. *et al.* Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nat. Methods* **14,** 600–606 (2017).

365. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **32,** 677–683 (2014).

366. Duan, J. *et al.* Genome-wide identification of CRISPR/Cas9 off-targets in human genome. *Cell Res.* **24,** 1009–1012 (2014).

367. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31,** 839–843 (2013).

368. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31,** 827–32 (2013).

369. Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529,** 490–495 (2016).

370. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. **351,** 84–89 (2016).

371. Chen, J. S. *et al.* Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550,** 407–410 (2017).

372. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32,** 279–284 (2014).

373. DeWitt, M. A. *et al.* Selection-free genome editing of the sickle mutation in human adult hematopoietic stem/progenitor cells. *Sci. Transl. Med.* **8,** 360ra134-360ra134 (2016).

374. Ye, L. *et al.* Genome editing using CRISPR-Cas9 to create the HPFH genotype in HSPCs: An approach for treating sickle cell disease and β-thalassemia. *Proc. Natl. Acad. Sci.* **113,** 10661–10665 (2016).

375. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39,** 1181 (2007).

376. Sheehan, V. A. *et al.* Whole exome sequencing identifies novel genes for fetal hemoglobin response to hydroxyurea in children with sickle cell anemia. *PLoS One* **9,** e110740 (2014).

377. Jones, E., Oliphant, E. & Peterson, P. SciPy: Open Source Scientific Tools for Python. (2001). Available at: http://www.scipy.org/.

378. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).

379. Gaudet, P. *et al.* The neXtProt knowledgebase on human proteins: Current status. *Nucleic Acids Res.* **43,** D764–D770 (2015).

380. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

381. Desktop Genetics. DESKGEN CLOUD. www.deskgen.com (2016).

382. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9–

mediated gene inactivation. *Nat. Biotechnol.* **32,** 1262–7 (2014).

383. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34,** 1–12 (2016).

384. Bialk, P., Rivera-Torres, N., Strouse, B. & Kmiec, E. B. Regulation of gene editing activity directed by single-stranded oligonucleotides and CRISPR/Cas9 systems. *PLoS One* **10,** 1–19 (2015).

385. Untergasser, A. *et al.* Primer3 - new capabilities and interfaces. *Nucleic Acids Res.* **40,** e115 (2012).

386. New England Biolabs. NEBaseChanger. Available at: http://nebasechanger.neb.com/.

387. GSL Biotech. SnapGene Viewer. Snapgene.com

388. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

389. EMBL-EBI. MUSCLE. Available at: www.ebi.ac.uk/Tools/msa/muscle/.

390. Lozzio, C. B. & Lozzio, B. B. Human Chronic Myelogenous Leukemia Cell-Line With Positive Philadelphia Chromosome. *Blood* **45,** 321–334 (1975).

391. Thoren, L. A. *et al.* Kit Regulates Maintenance of Quiescent Hematopoietic Stem Cells. *J. Immunol.* **180,** 2045–2053 (2008).

392. Edling, C. E. & Hallberg, B. c-Kit-A hematopoietic cell essential receptor tyrosine kinase. *Int. J. Biochem. Cell Biol.* **39,** 1995–1998 (2007).

393. Vitelli, L. *et al.* A pentamer transcriptional complex including tal-1 and retinoblastoma protein downmodulates c-kit expression in normal erythroblasts. *Mol. Cell. Biol.* **20,** 5330–5342 (2000).

394. Paulson, R. F., Shi, L. & Wu, D.-C. Stress erythropoiesis: new signals and new stress progenitor cells. *Curr. Opin. Hematol.* **18,** 139–45 (2011).

395. Illumina. Infinium ® HumanMethylation450 BeadChip. (2012).

396. Illumina. TruSeq Stranded mRNA Sample Preparation Guide. (2013). doi:# RS-122-9004DOC

397. Qiagen. AllPrep® DNA/RNA/Protein Mini Handbook. (2014).

398. Mollet, M., Godoy-Silva, R., Berdugo, C. & Chalmers, J. J. Computer simulations of the energy dissipation rate in a fluorescence-activated cell sorter: Implications to cells. *Biotechnol. Bioeng.* **100,** 260–272 (2008).

399. McKinney-Freeman, S. L. *et al.* Surface antigen phenotypes of hematopoietic stem cells

from embryos and murine embryonic stem cells. *Blood* **114,** 268–278 (2009).

400. Green, N. S. & Barral, S. Emerging science of hydroxyurea therapy for pediatric sickle cell disease. *Pediatr. Res.* **75,** 196–204 (2014).

401. Griffiths, R. E. *et al.* Maturing reticulocytes internalize plasma membrane in glycophorin A-containing vesicles that fuse with autophagosomes before exocytosis. *Blood* **119,** 6296–6306 (2012).

402. Wada, H. *et al.* Expression of major blood group antigens on human erythroid cells in a two phase liquid culture system. *Blood* **75,** 505–511 (1990).

403. Byrne, S. L. *et al.* Effect of glycosylation on the function of a soluble, recombinant form of the transferrin receptor. *Biochemistry* **45,** 6663–6673 (2006).

404. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 19096–101 (2009).

405. Lacey, S., Chung, J. Y. & Lin, H. A comparison of whole genome sequencing with exome sequencing for family-based association studies. *BMC Proc.* **8,** S38 (2014).

406. Kinney, T. *et al.* Silent cerebral infarcts in sickle cell anaemia: a risk factor analysis. *Paediatrics* **103,** 640–645 (1999).

407. Ohene-Frempong, K. *et al.* Cerebrovascular accidents in sickle cell disease: rates and risk factors. *Blood* **91,** 288–294 (1998).

408. Verduzco, L. a & Nathan, D. G. Review article Sickle cell disease and stroke. *Stroke* **114,** 5117–5125 (2009).

409. Bath, P. M. W. & Lees, K. R. Acute Stroke. *West. J. Med.* **173,** 209–212 (2000).

410. Heeney, M. M. & Ware, R. E. Hydroxyurea for children with sickle cell disease. *Pediatr. Clin. North Am.* **55,** 483–501, x (2008).

411. Jian, X., Boerwinkle, E. & Liu, X. In silico tools for splicing defect prediction - A survey from the viewpoint of end-users. *Genet. Med.* **16,** 497–503 (2014).

412. Burset, M., Seledtsov, I. A. & Solovyev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28,** 4364–4375 (2000).

413. Buratti, E. *et al.* Aberrant 5' splice sites in human disease genes: Mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res.* **35,** 4250–4263 (2007).

414. Vořechovský, I. Aberrant 3′ splice sites in human disease genes: Mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization.

Nucleic Acids Res. **34,** 4630–4641 (2006).

415. Kurmangaliyev, Y. Z., Sutormin, R. A., Naumenko, S. A., Bazykin, G. A. & Gelfand, M. S. Functional implications of splicing polymorphisms in the human genome. *Hum. Mol. Genet.* **22,** 3449–3459 (2013).

416. Lu, Z. X., Jiang, P. & Xing, Y. Genetic variation of pre-mRNA alternative splicing in human populations. *Wiley Interdiscip Rev RNA* **3,** 581–592 (2012).

417. Krawczak, M. *et al.* Single Base-Pair Substitutions in Exon–Intron Junctions of Human Genes: Nature, Distribution, and Consequences for mRNA Splicing. *Hum. Mutat.* **28,** 250–158 (2007).

418. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).

419. Thai, P., Loukoianov, A., Wachi, S. & Wu, R. Regulation of airway mucin gene expression. *Annu. Rev. Physiol.* **70,** 405–29 (2008).

420. Fuentes Fajardo, K. V *et al.* Detecting false-positive signals in exome sequencing. *Hum. Mutat.* **33,** 609–13 (2012).

421. Shyr, C. *et al.* FLAGS, frequently mutated genes in public exomes. *BMC Med. Genomics* **7,** 64 (2014).

422. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16,** 22 (2015).

423. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

424. Wu, Y. H. *et al.* Glucose-6-phosphate dehydrogenase enhances antiviral response through downregulation of NADPH Sensor HSCARG and upregulation of NF-κB signaling. *Viruses* **7,** 6689–6706 (2015).

425. Zhao, Y. *et al.* An NADPH sensor protein (HSCARG) down-regulates nitric oxide synthesis by association with argininosuccinate synthetase and is essential for epithelial cell viability. *J. Biol. Chem.* **283,** 11004–11013 (2008).

426. Maquat, L. E. Skiing Toward Nonstop mRNA Decay. *Science (80-.).* **295,** 2221–2222 (2002).

427. Myers, A. L. *et al.* IGFBP2 modulates the chemoresistant phenotype in esophageal adenocarcinoma. *Oncotarget* **6,** 25897–25916 (2015).

428. Gershtein, E. S. *et al.* Insulin-Like Growth Factors (IGF) and IGF-Binding Proteins

(IGFBP) in the Serum of Patients with Ovarian Tumors. *Bull. Exp. Biol. Med.* **160,** 814–816 (2016).

429. Huynh, H. *et al.* IGF binding protein 2 supports the survival and cycling of hematopoietic stem cells. *Blood* **118,** 3236–3243 (2011).

430. Kobayashi, H. *et al.* Angiocrine factors from Akt-activated endothelial cells balance self-renewal and differentiation of haematopietic stem cells. *Nat. Cell Biol.* **12,** 1046–1056 (2010).

431. Fan, D. M., Feng, X. S., Qi, P. W. & Chen, Y. W. Forkhead factor FOXQ1 promotes TGF-b1 expression and induces epithelial-mesenchymal transition. *Mol. Cell. Biochem.* **397,** 179–186 (2014).

432. Peng, X. *et al.* FOXQ1 mediates the crosstalk between TGF-b and Wnt signaling pathways in the progression of colorectal cancer. *Cancer Biol. Ther.* **16,** 1099–1109 (2015).

433. Christensen, J., Bentz, S., Sengstag, T., Shastri, V. P. & Anderle, P. FOXQ1, a Novel Target of the Wnt Pathway and a New Marker for Activation of Wnt Signaling in Solid Tumors. *PLoS One* **8,** 1–10 (2013).

434. Li, B. *et al.* Identification of Proteins Involved In Globin Gene Expression During Erythroid Maturation Using An In Vitro Erythroid Liquid Culture System. *Blood* **116,** 2065 (2010).

435. Zhang, X. *et al.* Inhibition of FOXQ1 induces apoptosis and suppresses proliferation in prostate cancer cells by controlling BCL11A/MDM2 expression. *Oncol. Rep.* **36,** 2349–2356 (2016).

436. Small, D. *et al.* STK-1, the human homolog of Flk-2/Flt-3, is selectively expressed in CD34+ human bone marrow cells and is involved in the proliferation of early progenitor/stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 459–63 (1994).

437. Lyman, S. D. *et al.* Molecular cloning of a ligand for the flt3/flk-2 tyrosine kinase receptor: A proliferative factor for primitive hematopoietic cells. *Cell* **75,** 1157–1167 (1993).

438. Janke, H. *et al.* Activating FLT3 mutants show distinct gain-of-function phenotypes in vitro and a characteristic signaling pathway profile associated with prognosis in acute Myeloid Leukemia. *PLoS One* **9,** e89560 (2014).

439. Felinski, E. A. & Quinn, P. G. The coactivator dTAF(II)110/hTAF(II)135 is sufficient to recruit a polymerase complex and activate basal transcription mediated by CREB. *Proc.*

Natl. Acad. Sci. U. S. A. **98,** 13078–13083 (2001).

440. Sangerman, J. *et al.* Mechanism for fetal hemoglobin induction by histone deacetylase inhibitors involves gamma-globin activation by CREB1 and ATF-2. *Blood* **108,** 3590–3599 (2006).

441. Islas, J. F. *et al.* Transcription factors ETS2 and MESP1 transdifferentiate human dermal fibroblasts into cardiac progenitors. *Proc. Natl. Acad. Sci.* **109,** 13016–13021 (2012).

442. Petrovic, N., Bhagwat, S. V., Ratzan, W. J., Ostrowski, M. C. & Shapiro, L. H. CD13/APN transcription is induced by RAS/MAPK-mediated phosphorylation of Ets-2 in activated endothelial cells. *J. Biol. Chem.* **278,** 49358–49368 (2003).

443. Ge, Y. *et al.* The role of the proto-oncogene ETS2 in acute megakaryocytic leukemia biology and therapy. *Leukemia* **22,** 521–9 (2008).

444. Tripathi, V. *et al.* Long Noncoding RNA MALAT1 Controls Cell Cycle Progression by Regulating the Expression of Oncogenic Transcription Factor B-MYB. *PLoS Genet.* **9,** e1003368 (2013).

445. Tripathi, V. *et al.* The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. *Mol. Cell* **39,** 925–938 (2010).

446. Baker, S. J. *et al.* B-myb is an essential regulator of hematopoietic stem cell and myeloid progenitor cell development. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 3122–3127 (2014).

447. Ma, X. Y. *et al.* Malat1 as an evolutionarily conserved lncRNA, plays a positive role in regulating proliferation and maintaining undifferentiated status of early-stage hematopoietic cells. *BMC Genomics* **16,** 676 (2015).

448. Geisler, S. & Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* **14,** 669–712 (2013).

449. Clegg, J. B. Can the product of the theta gene be a real globin? *Nature* **329,** 465–466 (1987).

450. Hsu, S.-L. *et al.* Structure and expression of the human theta1 globin gene. *Nature* **331,** 94–96 (1988).

451. Kanduri, C. Kcnq1ot1: A chromatin regulatory RNA. *Semin. Cell Dev. Biol.* **22,** 343–350 (2011).

452. Saxena, A. & Carninci, P. Long non-coding RNA modifies chromatin: Epigenetic silencing by long non-coding RNAs. *BioEssays* **33,** 830–839 (2011).

453. Tierling, S. *et al.* DNA methylation studies on imprinted loci in a male monozygotic twin pair discordant for Beckwith-Wiedemann syndrome. *Clin. Genet.* **79,** 546–553 (2011).

454. Chiesa, N. *et al.* The KCNQ1OT1 imprinting control region and non-coding RNA: New properties derived from the study of Beckwith-Wiedemann syndrome and Silver-Russell syndrome cases. *Hum. Mol. Genet.* **21,** 10–25 (2012).

455. Dunn, O. J. Multiple Comparisons Among Means. *J. Am. Stat. Assoc.* **56,** 52–64 (1961).

456. Lacey, S., Chung, J. Y. & Lin, H. A comparison of whole genome sequencing with exome sequencing for family-based association studies. *BMC Proc.* **8,** S38 (2014).

457. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet.* **24,** 1202–1205 (2016).

458. Chakalova, L. *et al.* The Corfu delta/beta thalassemia deletion disrupts gamma-globin gene silencing and reveals post-transcriptional regulation of HbF expression. *Blood* **105,** 2154–2160 (2005).

459. Preumont, A., Rzem, R., Vertommen, D. & Van Schaftingen, E. HDHD1, which is often deleted in X-linked ichthyosis, encodes a pseudouridine-5'-phosphatase. *Biochem. J.* **431,** 237–44 (2010).

460. Chen, H. *et al.* Decreased hephaestin expression and activity leads to decreased iron efflux from differentiated Caco2 cells. *J. Cell. Biochem.* **107,** 803–808 (2009).

461. Denninger, J. W. & Marletta, M. A. Guanylate cyclase and the NO/cGMP signaling pathway. *Biochim. Biophys. Acta* **1411,** 334–350 (1999).

462. The International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478,** 103–109 (2011).

463. Lou, T.-F., Singh, M., Mackie, A., Li, W. & Pace, B. S. Hydroxyurea generates nitric oxide in human erythroid cells: mechanisms for gamma-globin gene activation. *Exp. Biol. Med. (Maywood).* **234,** 1374–82 (2009).

464. Aveic, S., Pigazzi, M. & Basso, G. BAG1: The guardian of anti-apoptotic proteins in acute myeloid leukemia. *PLoS One* **6,** (2011).

465. Götz, R. *et al.* Essential role of Bag-1 in differentiation and survival of hematopoietic and neuronal cells Rudolf. *Nat. Neurosci.* **8,** 1169–1178 (2005).

466. Korf-Klingebiel, M. *et al.* Myeloid-derived growth factor (C19orf10) mediates cardiac

repair following myocardial infarction. *Nat. Med.* **21,** 140–149 (2015).

467. Paquet, D. *et al.* Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533,** 125–129 (2016).

468. GenScript. GenScript Codon Usage Frequence Table Tool.

469. Chen, F. *et al.* High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat. Methods* **8,** 753–5 (2011).

470. Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L. & Corn, J. E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat. Biotechnol.* **34,** 339–344 (2016).

471. Lohmann, F. & Bieker, J. J. Activation of Eklf expression during hematopoiesis by Gata2 and Smad5 prior to erythroid commitment. *Development* **135,** 2071–2082 (2008).

472. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

473. Wang, J. *et al.* Factorbook.org: A Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.* **41,** 171–176 (2012).

474. Rosenbloom, K. R. *et al.* ENCODE Data in the UCSC Genome Browser: Year 5 update. *Nucleic Acids Res.* **41,** 56–63 (2013).

475. Thein, S. L., Wood, W. G., Wickramasinghe, S. N. & Galvin, M. C. Beta-Thalassemia Unlinked to the Beta-Globin Gene in an English Family. *Blood* **82,** 961–967 (1993).

476. Breton, A. *et al.* ASH1L (a histone methyltransferase protein) is a novel candidate globin gene regulator revealed by genetic study of an English family with beta-thalassaemia unlinked to the beta-globin locus. *Br. J. Haematol.* **175,** 525–530 (2016).

477. Tanaka, Y., Katagiri, Z. ichiro, Kawahashi, K., Kioussis, D. & Kitajima, S. Trithorax-group protein ASH1 methylates histone H3 lysine 36. *Gene* **397,** 161–168 (2007).

478. Gregory, G. D. *et al.* Mammalian ASH1L Is a Histone Methyltransferase That Occupies the Transcribed Region of Active Genes. *Mol. Cell. Biol.* **27,** 8466–8479 (2007).

479. Yuan, W. *et al.* H3K36 methylation antagonizes PRC2-mediated H3K27 methylation. *J. Biol. Chem.* **286,** 7983–7989 (2011).

480. Breton, A. *et al.* ASH1L: A Novel Beta-Globin Gene Regulator in Humans. *Blood* **126,** 641 (2015).

481. Nakamura, T. *et al.* huASH1 protein, a putative transcription factor encoded by a human homologue of the Drosophila ash1 gene, localizes to both nuclei and cell-cell tight

junctions. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 7284–9 (2000).

482. Aravind, L. & Landsman, D. AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Res.* **26,** 4413–4421 (1998).

483. Sanchez, R., Meslamani, J. & Zhou, M.-M. The Bromodomain: From Epigenome Reader to Druggable Target. *Biochim. Biophys. Acta* (2014). doi:10.1016/j.bbagrm.2014.03.011

484. Sanchez, R. & Zhou, M. M. The PHD finger: A versatile epigenome reader. *Trends Biochem. Sci.* **36,** 364–372 (2011).

485. Li, H. *et al.* Efficient CRISPR-Cas9 mediated gene disruption in. *Haematologica* **101,** e216–e219 (2016).

486. Varga, E., Hansen, M., Wüst, T., von Lindern, M. & van den Akker, E. Generation of human erythroblast-derived iPSC line using episomal reprogramming system. *Stem Cell Res.* **25,** 30–33 (2017).

487. Yang, C. T. *et al.* Human induced pluripotent stem cell derived erythroblasts can undergo definitive erythropoiesis and co-express gamma and beta globins. *Br. J. Haematol.* **166,** 435–448 (2014).

488. Niu, X. *et al.* Combining single strand oligodeoxynucleotides and CRISPR/Cas9 to correct gene mutations in β-thalassemia-induced pluripotent stem sells. *J. Biol. Chem.* **291,** 16576–16585 (2016).

489. Kurita, R. *et al.* Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. *PLoS One* **8,** e59890 (2013).

490. Trakarnsanga, K. *et al.* An immortalized adult human erythroid line facilitates sustainable and scalable generation of functional red cells. *Nat. Commun.* **8,** (2017).

491. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527,** 192–197 (2015).

492. Yien, Y. Y. & Bieker, J. J. Functional interactions between erythroid kruppel-like factor (EKLF/KLF1) and protein phosphatase PPM1B/PP2C?? *J. Biol. Chem.* **287,** 15193–15204 (2012).

493. Villamizar, O. *et al.* Data in support of transcriptional regulation and function of Fas-antisense long noncoding RNA during human erythropoiesis. *Data Br.* **7,** 1288–1295 (2016).

494. Zhang, D., Cho, E. & Wong, J. A critical role for the co-repressor N-CoR in erythroid differentiation and heme synthesis. *Cell Res.* **17,** 804–814 (2007).

495. Naumann, S., Reutzel, D., Speicher, M. & Decker, H. J. Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. *Leuk. Res.* **25,** 313–322 (2001).

496. Komatsu, K., Nakamura, H., Shinkai, K. & Akedo, H. Secretion of Transforming Growth Factor-Beta by Human Myelogenous Leukemic Cells and Its Possible Role in Proliferation of the Leukemic Cells. *Japanese J. Cancer Res.* **80,** 928–931 (1989).

497. He, R. *et al.* Inhibition of K562 leukemia angiogenesis and growth by expression of antisense vascular endothelial growth factor (VEGF) sequence. *Cancer Gene Ther.* **10,** 879–886 (2003).

498. Mari, P.-O. *et al.* Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4. *Pnas* **103,** 18597–18602 (2006).

499. Zhang, Z. *et al.* Solution structure of the C-terminal domain of Ku80 suggests important sites for protein-protein interactions. *Structure* **12,** 495–502 (2004).

500. Mimori, T. & Hardin, J. A. Mechanism of interaction between Ku protein and DNA. *J. Biol. Chem.* **261,** 10375–10379 (1986).

501. Fattah, F. *et al.* Ku Regulates the Non-Homologous End Joining Pathway Choice of DNA Double-Strand Break Repair in Human Somatic Cells. *PLoS Genet.* **6,** e1000855 (2010).

502. Davis, A. J. & Chen, D. J. DNA double strand break repair via non-homologous end-joining. *Transl. Cancer Res.* **2,** 130–43 (2013).

503. Grawunder, U. *et al.* Activity of DNA ligase IV stimulated by complex formation with XRCC4 protein in mammalian cells. *Nature* **388,** 492–5 (1997).

504. Wei, P. C., Lo, W. T., Su, M. I., Shew, J. Y. & Lee, W. H. Non-targeting siRNA induces NPGPx expression to cooperate with exoribonuclease XRN2 for releasing the stress. *Nucleic Acids Res.* **40,** 323–332 (2012).

505. N Calculators. Poisson Distribution Formula & Calculations. Available at: http://ncalculators.com/statistics/poisson-distribution-calculator.htm. (Accessed: 1st January 2017)

506. Addya, S. *et al.* Erythroid-induced commitment of K562 cells results in clusters of differentially expressed genes enriched for specific transcription regulatory elements. *Physiol. Genomics* **19,** 117–130 (2004).

507. Witt, O., Schulze, S., Kanbach, K., Roth, C. & Pekrun, a. Tumor cell differentiation by

butyrate and environmental stress. *Cancer Lett.* **171,** 173–182 (2001).

508. Chen, K. *et al.* Resolving the distinct stages in erythroid differentiation based on dynamic changes in membrane protein expression during erythropoiesis. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 17413–8 (2009).

509. Fleige, S. & Pfaffl, M. W. RNA integrity and the effect on the real-time qRT-PCR performance. *Mol. Aspects Med.* **27,** 126–139 (2006).

510. Parkins, A. C., Sharpe, A. H. & Orkin, S. H. Lethal β-thalassaemia in mice lacking the erythroid CACCC-transcription factor EKLF. *Nature* **375,** 318–322 (1995).

511. Drissen, R. *et al.* The active spatial organization of the β-globin locus requires the transcription factor EKLF. *Genes Dev.* **18,** 2485–2490 (2004).

512. Tallack, M. R. *et al.* A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res.* **20,** 1052–1063 (2010).

513. Bhanu, N. V. *et al.* A sustained and pancellular reversal of gamma-globin gene silencing in adult human erythroid precursor cells. *Blood* **105,** 387–393 (2005).

514. Xiang, J., Wu, D. C., Chen, Y. & Paulson, R. F. In vitro culture of stress erythroid progenitors identifies distinct progenitor populations and analogous human progenitors. *Blood* **125,** 1803–1812 (2015).

515. Borrione, P. *et al.* A biparametric flow cytometry analysis for the study of reticulocyte patterns of maturation. *Int. J. Lab. Hematol.* **32,** 65–73 (2010).

516. Kaushal, M. *et al.* Examination of Reticulocytosis among Chronically Transfused Children with Sickle Cell Anemia. *PLoS One* **11,** e0153244 (2016).

517. Sato, S. *et al.* Enhanced expression of CD71, transferrin receptor, on immature reticulocytes in patients with paroxysmal nocturnal hemoglobinuria. *Int. J. Lab. Hematol.* **32,** 137–143 (2010).

518. Malleret, B. *et al.* Significant Biochemical, Biophysical and Metabolic Diversity in Circulating Human Cord Blood Reticulocytes. *PLoS One* **8,** e76062 (2013).

519. Serke, S. & Huhn, D. Identification of CD71 (transferrin receptor) expressing erythrocytes by multiparameter-flow-cytometry (MP-FCM): correlation to the quantitation of reticulocytes as determined by conventional microscopy and by MP-FCM using a RNA-staining dye. *Br. J. Haematol.* **81,** 432–439 (1992).

520. Luck, L., Zeng, L., Hiti, A. L., Weinberg, K. I. & Malik, P. Human CD34+ and CD34+CD38- hematopoietic progenitors in sickle cell disease differ phenotypically and

functionally from normal and suggest distinct subpopulations that generate F cells. *Exp. Hematol.* **32,** 483–493 (2004).

521. Tishkoff, S. A. & Williams, S. M. Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* **3,** 611–621 (2002).

522. Reiner, A. P. *et al.* Population structure, admixture, and aging-related phenotypes in African American adults: the Cardiovascular Health Study. *Am J Hum Genet* **76,** 463–477 (2005).

523. Misko, T. P., Schilling, R. J., Salvemini, D., Moore, W. M. & Currie, M. G. A Fluorometric Assay for the Measurement of Nitrate in Biological Samples. *Anal. Biochem.* **214,** 11–16 (1993).

524. Pászty, C. *et al.* Transgenic knockout mice with exclusively human sickle hemoglobin and sickle cell disease. *Science* **278,** 876–878 (1997).

525. Chua, C. Y. *et al.* IGFBP2 potentiates nuclear EGFR-STAT3 signaling. *Oncogene* **35,** 738–747 (2016).

526. Maroulakou, I. G. & Bowe, D. B. Expression and function of Ets transcription factors in mammalian development: a regulatory network. *Oncogene* **19,** 6432–6442 (2000).

527. Knee, D. A., Froesch, B. A., Nuber, U., Takayama, S. & Reed, J. C. Structure-function analysis of Bag1 proteins. Effects on androgen receptor transcriptional activity. *J. Biol. Chem.* **276,** 12718–12724 (2001).

528. McLane, L. M. & Corbett, A. H. Nuclear localization signals and human disease. *IUBMB Life* **61,** 697–706 (2009).

529. Takayama, S. *et al.* Cloning and functional analysis of BAG-1: A novel Bcl-2-binding protein with anti-cell death activity. *Cell* **80,** 279–284 (1995).

530. Chen, J., Xiong, J., Liu, H., Chernenko, G. & Tang, S.-C. Distinct BAG-1 isoforms have different anti-apoptotic functions in BAG-1-transfected C33A human cervical carcinoma cell line. *Oncogene* **21,** 7050–7059 (2002).

531. Domen, J., Cheshier, S. H. & Weissman, I. L. The Role of Apoptosis in the Regulation of Hematopoietic Stem Cells: Overexpression of BCL-2 Increases Both Their Number and Repopulation Potential. *J. Exp. Med.* **191,** 253–264 (2000).

532. Haughn, L., Hawley, R. G., Morrison, D. K., Von Boehmer, H. & Hockenbery, D. M. BCL-2 and BCL-XL Restrict Lineage Choice during Hematopoietic Differentiation. *J. Biol. Chem.* **278,** 25158–25165 (2003).

533. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

534. Borg, J., Patrinos, G. P., Felice, A. E. & Philipsen, S. Erythroid phenotypes associated with KLF1 mutations. *Haematologica* **96,** 635–638 (2011).

535. Natiq, A. *et al.* Hereditary persistence of fetal hemoglobin in two patients with KLF1 haploinsufficiency due to 19p13.2–p13.12/13 deletion. *Am. J. Hematol.* **92,** E2–E3 (2017).

536. DeBaun, M. R. *et al.* Controlled Trial of Transfusions for Silent Cerebral Infarcts in Sickle Cell Anemia. *N. Engl. J. Med.* **371,** 699–710 (2014).

537. Quinn, C. T. *et al.* Acute Silent Cerebral Ischemic Events in Children with Sickle Cell Anaemia. *JAMA Neurol.* **70,** 58–65 (2013).

538. Miller, S. T. *et al.* Silent infarction as a risk factor for overt stroke in children with sickle cell anemia: A report from the Cooperative Study of Sickle Cell Disease. *J. Pediatr.* **139,** 385–390 (2001).

539. Kim, J. A. *et al.* A novel electroporation method using a capillary and wire-type electrode. *Biosens. Bioelectron.* **23,** 1353–1360 (2008).

540. Lonza. Cell and Transfection Database.

541. Brinkman, E. K., Chen, T., Amendola, M. & Van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42,** e168 (2014).

542. Kim, J. M., Kim, D., Kim, S. & Kim, J. S. Genotyping with CRISPR-Cas-derived RNA-guided endonucleases. *Nat. Commun.* **5,** 1–7 (2014).

543. Altmann, T. & Gennery, A. R. DNA ligase IV syndrome; a review. *Orphanet J. Rare Dis.* **11,** 137 (2016).

544. Wang, B. *et al.* Role of Ku70 in deubiquitination of Mcl-1 and suppression of apoptosis. *Cell Death Differ.* **21,** 1160–1169 (2014).

545. Delgado-Cañedo, A., Santos, D. G. Dos, Chies, J. A. B., Kvitko, K. & Nardi, N. B. Optimization of an electroporation protocol using the K562 cell line as a model: Role of cell cycle phase and cytoplasmic DNAses. *Cytotechnology* **51,** 141–148 (2006).

546. Yoshimi, K. *et al.* ssODN-mediated knock-in with CRISPR-Cas for large genomic regions in zygotes. *Nat. Commun.* **7,** 10431 (2016).

547. Renaud, J. B. *et al.* Improved Genome Editing Efficiency and Flexibility Using Modified

Oligonucleotides with TALEN and CRISPR-Cas9 Nucleases. *Cell Rep.* **14,** 2263–2272 (2016).

548. Liang, X. *et al.* Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. *J. Biotechnol.* **208,** 44–53 (2015).

549. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24,** 1012–1019 (2014).

550. Mabaera, R. *et al.* Developmental- and differentiation-specific patterns of human γ- and β-globin promoter DNA methylation. *Hematology* **110,** 3–5 (2007).

551. Kim, A., Kiefer, C. M. & Dean, A. Distinctive Signatures of Histone Methylation in Transcribed Coding and Noncoding Human -Globin Sequences. *Mol. Cell. Biol.* **27,** 1271–1279 (2007).

552. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

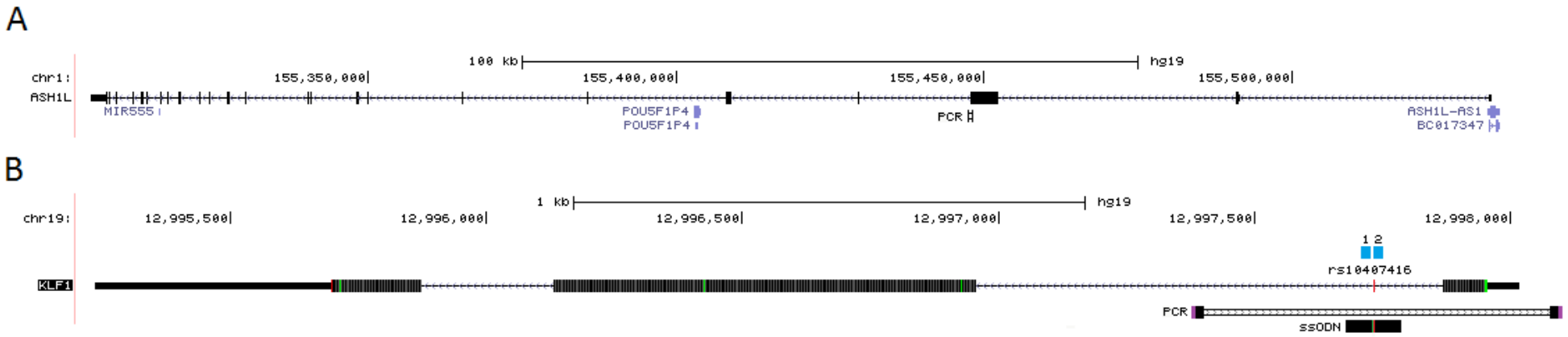
553. Trakarnsanga, K. *et al.* Induction of adult levels of β-globin in human erythroid cells that intrinsically express embryonic or fetal globin by transduction with KLF1 and BCL11A-XL. *Haematologica* **99,** 1677–1685 (2014).

554. Yien, Y. Y. & Bieker, J. J. EKLF/KLF1, a Tissue-Restricted Integrator of Transcriptional Control, Chromatin Remodeling, and Lineage Determination. *Mol. Cell. Biol.* **33,** 4–13 (2013).

555. Siatecka, M., Soni, S., Planutis, A. & Bieker, J. J. Transcriptional activity of erythroid kruppel-like factor (EKLF/KLF1) modulated by PIAS3 (Protein inhibitor of activated STAT3). *J. Biol. Chem.* **290,** 9929–9940 (2015).

Appendix 1

Full gene maps of the ASH1L and KLF1 genes (A & B respectively), adapted from the UCSC Genome Browser (*http://genome.ucsc.edu* - Assembly GRCh37/hg19[380]). PCR amplicons for template regions amplified for cloning into CRISPR-Cas9 plasmids are annotated. ASH1L gene spans a much larger area, and contains many exons. Structure of the KLF1 gene is much simpler, with only three exons, the location of the intronic SNP is also highlighted in red.



Appendix 2

CRISPR-Cas9 Plasmid Sequence. 20bp variable gRNA target sequence is highlighted in blue, with the conserved regions of the gRNA SDM primers highlighted in yellow. The BssHII restriction enzyme site used for template DNA insertion is shown in magenta. Cas9 sequence is shown in red in bold, and the GFP sequence is highlighted in green.

CCGNNNNNNNNNNNNNNNNNNGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCC
GTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTGAAGACACACCGCTCTCAATATTGGCC
ATTAGCCATATTATTCATTGGTTATATAGCATAAATCAATATTGGCTATTGGCCATTGCATACGTTGTAT
CTATATCATAATATGTACATTTATATTGGCTCATGTCCAATATGACCGCCATGTTGGCATTGATTATTGA
CTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACAT
AACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACG
TATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAAC
TGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTCCGCCCCCTATTGACGTCAATGACGGTA
AATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTACGGGACTTTCCTACTTGGCAGTACATCTACG
TATTAGTCATCGCTATTACCATGCTGATGCGGTTTTGGCAGTACACCAATGGGCGTGGATAGCGGTTT
GACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCA
ACGGGACTTTCCAAAATGTCGTAATAACCCCGCCCCGTTGACGCAAATGGGCGGTAGGCGTGTACGG
TGGGAGGTCTATATAAGCAGAGCTCGTTTAGTGAACCGTCAGATCACTAGAAGCTTTATTGCGGTAGT
TTATCACAGTTAAATTGCTAACGCAGTCAGGCCAACAGAGACCACACCCAAGCTGGCCGCCACCATGG
CCCCAAAGAAGAAGCGGAAGGTCGGTATCCACGGAGTCCCAGCAGCC**GACAAGAAGTACAGCATCG**
GCCTGGACATCGGCACCAACTCTGTGGGCTGGGCCGTGATCACCGACGAGTACAAGGTGCCCAGCA
AGAAATTCAAGGTGCTGGGCAACACCGACCGGCACAGCATCAAGAAGAACCTGATCGGAGCCCTG
CTGTTCGACAGCGGCGAAACAGCCGAGGCCACCCGGCTGAAGAGAACCGCCAGAAGAAGATACAC
CAGACGGAAGAACCGGATCTGCTATCTGCAAGAGATCTTCAGCAACGAGATGGCCAAGGTGGACG
ACAGCTTCTTCCACAGACTGGAAGAGTCCTTCCTGGTGGAAGAGGATAAGAAGCACGAGCGGCACC
CCATCTTCGGCAACATCGTGGACGAGGTGGCCTACCACGAGAAGTACCCCACCATCTACCACCTGAG
AAAGAAACTGGTGGACAGCACCGACAAGGCCGACCTGCGGCTGATCTATCTGGCCCTGGCCCACAT
GATCAAGTTCCGGGGCCACTTCCTGATCGAGGGCGACCTGAACCCCGACAACAGCGACGTGGACAA
GCTGTTCATCCAGCTGGTGCAGACCTACAACCAGCTGTTCGAGGAAAACCCCATCAACGCCAGCGGC
GTGGACGCCAAGGCCATCCTGTCTGCCAGACTGAGCAAGAGCAGACGGCTGGAAAATCTGATCGCC
CAGCTGCCCGGCGAGAAGAAGAATGGCCTGTTCGGAAACCTGATTGCCCTGAGCCTGGGCCTGACC
CCCAACTTCAAGAGCAACTTCGACCTGGCCGAGGATGCCAAACTGCAGCTGAGCAAGGACACCTAC
GACGACGACCTGGACAACCTGCTGGCCCAGATCGGCGACCAGTACGCCGACCTGTTTCTGGCCGCC
AAGAACCTGTCCGACGCCATCCTGCTGAGCGACATCCTGAGAGTGAACACCGAGATCACCAAGGCC
CCCCTGAGCGCCTCTATGATCAAGAGATACGACGAGCACCACCAGGACCTGACCCTGCTGAAAGCTC
TCGTGCGGCAGCAGCTGCCTGAGAAGTACAAAGAGATTTTCTTCGACCAGAGCAAGAACGGCTACG
CCGGCTACATTGACGGCGGAGCCAGCCAGGAAGAGTTCTACAAGTTCATCAAGCCCATCCTGGAAA
AGATGGACGGCACCGAGGAACTGCTCGTGAAGCTGAACAGAGAGGACCTGCTGCGGAAGCAGCG
GACCTTCGACAACGGCAGCATCCCCCACCAGATCCACCTGGGAGAGCTGCACGCCATTCTGCGGCG
GCAGGAAGATTTTTACCCATTCCTGAAGGACAACCGGGAAAAGATCGAGAAGATCCTGACCTTCCG
CATCCCCTACTACGTGGGCCCTCTGGCCAGGGGAAACAGCAGATTCGCCTGGATGACCAGAAAGAG
CGAGGAAACCATCACCCCCTGGAACTTCGAGGAAGTGGTGGACAAGGGCGCTTCCGCCCAGAGCTT
CATCGAGCGGATGACCAACTTCGATAAGAACCTGCCCAACGAGAAGGTGCTGCCCAAGCACAGCCT
GCTGTACGAGTACTTCACCGTGTATAACGAGCTGACCAAAGTGAAATACGTGACCGAGGGAATGAG
AAAGCCCGCCTTCCTGAGCGGCGAGCAGAAAAAGGCCATCGTGGACCTGCTGTTCAAGACCAACCG
GAAAGTGACCGTGAAGCAGCTGAAAGAGGACTACTTCAAGAAAATCGAGTGCTTCGACTCCGTGG
AAATCTCCGGCGTGGAAGATCGGTTCAACGCCTCCCTGGGCACATACCACGATCTGCTGAAAATTAT
CAAGGACAAGGACTTCCTGGACAATGAGGAAAACGAGGACATTCTGGAAGATATCGTGCTGACCCT
GACACTGTTTGAGGACAGAGAGATGATCGAGGAACGGCTGAAAACCTATGCCCACCTGTTCGACGA
CAAAGTGATGAAGCAGCTGAAGCGGCGGAGATACACCGGCTGGGGCAGGCTGAGCCGGAAGCTG
ATCAACGGCATCCGGGACAAGCAGTCCGGCAAGACAATCCTGGATTTCCTGAAGTCCGACGGCTTC

GCCAACAGAAACTTCATGCAGCTGATCCACGACGACAGCCTGACCTTTAAAGAGGACATCCAGAAA
GCCCAGGTGTCCGGCCAGGGCGATAGCCTGCACGAGCACATTGCCAATCTGGCCGGCAGCCCCGCC
ATTAAGAAGGGCATCCTGCAGACAGTGAAGGTGGTGGACGAGCTCGTGAAAGTGATGGGCCGGCA
CAAGCCCGAGAACATCGTGATCGAAATGGCCAGAGAGAACCAGACCACCCAGAAGGGACAGAAG
AACAGCCGCGAGAGAATGAAGCGGATCGAAGAGGGCATCAAAGAGCTGGGCAGCCAGATCCTGA
AAGAACACCCCGTGGAAAACACCCAGCTGCAGAACGAGAAGCTGTACCTGTACTACCTGCAGAATG
GGCGGGATATGTACGTGGACCAGGAACTGGACATCAACCGGCTGTCCGACTACGATGTGGACCATA
TCGTGCCTCAGAGCTTTCTGAAGGACGACTCCATCGACAACAAGGTGCTGACCAGAAGCGACAAGA
ACCGGGGCAAGAGCGACAACGTGCCCTCCGAAGAGGTCGTGAAGAAGATGAAGAACTACTGGCG
GCAGCTGCTGAACGCCAAGCTGATTACCCAGAGAAAGTTCGACAATCTGACCAAGGCCGAGAGAG
GCGGCCTGAGCGAACTGGATAAGGCCGGCTTCATCAAGAGACAGCTGGTGGAAACCCGGCAGATC
ACAAAGCACGTGGCACAGATCCTGGACTCCCGGATGAACACTAAGTACGACGAGAATGACAAGCT
GATCCGGGAAGTGAAAGTGATCACCCTGAAGTCCAAGCTGGTGTCCGATTTCCGGAAGGATTTCCA
GTTTTACAAAGTGCGCGAGATCAACAACTACCACCACGCCCACGACGCCTACCTGAACGCCGTCGTG
GGAACCGCCCTGATCAAAAAGTACCCTAAGCTGGAAAGCGAGTTCGTGTACGGCGACTACAAGGTG
TACGACGTGCGGAAGATGATCGCCAAGAGCGAGCAGGAAATCGGCAAGGCTACCGCCAAGTACTT
CTTCTACAGCAACATCATGAACTTTTTCAAGACCGAGATTACCCTGGCCAACGGCGAGATCCGGAAG
CGGCCTCTGATCGAGACAAACGGCGAAACCGGGGAGATCGTGTGGGATAAGGGCCGGGATTTTGC
CACCGTGCGGAAAGTGCTGAGCATGCCCCAAGTGAATATCGTGAAAAAGACCGAGGTGCAGACAG
GCGGCTTCAGCAAAGAGTCTATCCTGCCCAAGAGGAACAGCGATAAGCTGATCGCCAGAAAGAAG
GACTGGGACCCTAAGAAGTACGGCGGCTTCGACAGCCCCACCGTGGCCTATTCTGTGCTGGTGGTG
GCCAAAGTGGAAAAGGGCAAGTCCAAGAAACTGAAGAGTGTGAAAGAGCTGCTGGGGATCACCA
TCATGGAAAGAAGCAGCTTCGAGAAGAATCCCATCGACTTTCTGGAAGCCAAGGGCTACAAAGAA
GTGAAAAAGGACCTGATCATCAAGCTGCCTAAGTACTCCCTGTTCGAGCTGGAAAACGGCCGGAAG
AGAATGCTGGCCTCTGCCGGCGAACTGCAGAAGGGAAACGAACTGGCCCTGCCCTCCAAATATGTG
AACTTCCTGTACCTGGCCAGCCACTATGAGAAGCTGAAGGGCTCCCCCGAGGATAATGAGCAGAAA
CAGCTGTTTGTGGAACAGCACAAGCACTACCTGGACGAGATCATCGAGCAGATCAGCGAGTTCTCC
AAGAGAGTGATCCTGGCCGACGCTAATCTGGACAAAGTGCTGTCCGCCTACAACAAGCACCGGGAT
AAGCCCATCAGAGAGCAGGCCGAGAATATCATCCACCTGTTTACCCTGACCAATCTGGGAGCCCCTG
CCGCCTTCAAGTACTTTGACACCACCATCGACCGGAAGAGGTACACCAGCACCAAAGAGGTGCTGG
ACGCCACCCTGATCCACCAGAGCATCACCGGCCTGTACGAGACACGGATCGACCTGTCTCAGCTGGG
AGGCGACAAAAGGCCGGCGGCCACGAAAAAGGCCGGCCAGGCAAAAAAGAAAAAGGGCACATCTG
AGGGCAGGGGAAGTCTGCTAACATGCGGGGACGTGGAGGAAAATCCCGGCCCCATGACTGCCCTGA
CCGAAGGTGCTAAGCTGTTTGAGAAGGAGATTCCGTACATCACCGAGCTGGAAGGGGACGTCGAAG
GAATGAAGTTCATCATCAAGGGAGAAGGAACCGGGGACGCTACGACTGGAACCATTAAGGCCAAGT
ATATCTGTACCACTGGAGATCTGCCAGTGCCTTGGGCCACCCTTGTGTCAACCCTCTCGTATGGAGTGC
AGTGTTTTGCTAAGTACCCTAGCCACATTAAGGACTTCTTCAAATCCGCCATGCCGGAAGGTTATACCC
AAGAGCGCACCATTTCTTTTGAGGGAGATGGAGTGTACAAGACCCGCGCGATGGTCACCTATGAGAG
GGGATCTATCTACAACCGGGTGACTCTGACTGGAGAAAACTTTAAGAAGGACGGGCATATTCTTCGG
AAGAATGTCGCCTTCCAGTGCCCTCCCAGCATCCTTTACATTCTCCCCGACACTGTGAACAACGGAATC
CGCGTGGAGTTCAATCAAGCCTACGACATCGAGGGGGTGACGGAGAAGCTGGTGACCAAGTGTAGC
CAGATGAATCGGCCACTGGCCGGTTCAGCGGCTGTCCACATTCCGCGCTACCATCATATCACTTATCAC
ACTAAGCTCTCCAAAGACCGCGATGAGAGGAGAGATCACATGTGCCTGGTGGAAGTGGTCAAGGCC
GTCGATCTCGATACCTATCAGTAAGGTCTCACTCGAGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGC
CATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTA
ATAAAATGAGGAAATTGCATCCCCACTTCAGAAGTTCCTATACTTTCTAGAGAATAGGAACTTCACTAT
AGAGTCGAATAAGGGCGACACCCCCTAATTAGCCCGGGCGAAAGGCCCAGTCTTTCGACTGAGCCTT
TCGTTTTATTTGATGCCTGGCAGTTCCCTACTCTCGCATGGGGAGTCCCCACACTACCATCGGCGCTAC
GGCGTTTCACTTCTGAGTTCGGCATGGGGTCAGGTGGGACCACCGCGCTACTGCCGCCAGGCAAACA
AGGGGTGTTATGAGCCATATTCAGGTATAAATGGGCTCGCGATAATGTTCAGAATTGGTTAATTGGTT
GTAACACTGACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAAC
CCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAATATGAGCCATATTCAACGGGAAACGTCGAG
GCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGC
AATCAGGTGCGACAATCTATCGCTTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGC

AAAGGTAGCGTTGCCAATGATGTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGC
CACTTCCGACCATCAAGCATTTTATCCGTACTCCTGATGATGCATGGTTACTCACCACTGCGATCCCCG
GAAAAACAGCGTTCCAGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGGCA
GTGTTCCTGCGCCGGTTGCACTCGATTCCTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGCC
TCGCTCAGGCGCAATCACGAATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAAT
GGCTGGCCTGTTGAACAAGTCTGGAAAGAAATGCATAAACTTTTGCCATTCTCACCGGATTCAGTCGT
CACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTGATGT
TGGACGAGTCGGAATCGCAGACCGATACCAGGATCTTGCCATCCTATGGAACTGCCTCGGTGAGTTTT
CTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTATTGATAATCCTGATATGAATAAATTGCAGT
TTCATTTGATGCTCGATGAGTTTTTCTAAAAGCAGAGCATTACGCTGACTTGACGGGACGGCGCAAGC
TCATGACCAAAATCCCTTAACGTGAGTTACGCGCGCGTCGTTCCACTGAGCGTCAGACCCCGTAGAAA
AGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCAC
CGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTC
AGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGCCGTAGTTAGCCCACCACTTCAAGAACTC
TGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTC
GTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGG
GGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAG
CTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTC
GGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGG
TTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAA
CGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCG
TTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGA
ACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGGCGAGAGTAGGGAACTGCCAGGCAT
CAAACTAAGCAGAAGGCCCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTCTGTGTTGTAAAA
CGACGGCCAGTCTTAAGCTCGGGCCCCCTGGGCGGTTCTGATAACGAGTAATCGTTAATCCGCAAATA
ACGTAAAAACCCGCTTCGGCGGGTTTTTTTATGGGGGGAGTTTAGGGAAAGAGCATTTGTCAGAATA
TTTAAGGGCGCCTGTCACTTTGCTTGATATATGAGAATTATTTAACCTTATAAATGAGAAAAAAGCAAC
GCACTTTAAATAAGATACGTTGCTTTTTCGATTGATGAACACCTATAATTAAACTATTCATCTATTATTT
ATGATTTTTTGTATATACAATATTTCTAGTTTGTTAAAGAGAATTAAGAAAATAAATCTCGAAAATAAT
AAAGGGAAAATCAGTTTTTGATATCAAAATTATACATGTCAACGATAATACAAAATATAATACAAACT
ATAAGATGTTATCAGTATTTATTATGCATTTAGAATAAATTTTGTGTCGCCCTTATTCGACTCACTATAG
AAGTTCCTATTCTCTAGAAAGTATAGGAACTTCACTTCATTTTCCGTCTTCGAGGGCCTATTTCCCATGA
TTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGAGATAATTGGAATTAATTTGACTGTAAACA
CAAAGATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAA
TTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATA
TCTTGTGGAAAGGACGAAACA

Appendix 3

List of SDM primers used to substitute gRNA sequences into CRISPR-Cas9 plasmids. 10bp variable sequence at 5' of each plasmid is shown in bold in red.

Primer	*Sequence*
ASH1L_gRNA_1F_b	**ACTGGAGTTA**GTTTTAGAGCTAGAAATAGCAAG
ASH1L_gRNA1_SDMR	**GGCCGGAAGA**CGGTGTTTCGTCCTTTCC
ASH1L_gRNA_2F_b	**CTCCAGTGGC**GTTTTAGAGCTAGAAATAGCAAG
ASH1L_gRNA2_SDMR	**TTAGGGTTTG**CGGTGTTTCGTCCTTTCC
KLF1_gRNA_1F_b	**TAGTCTGGCA**GTTTTAGAGCTAGAAATAGCAAG
KLF1_gRNA1_SDMR	**AGCTGAGATC**CGGTGTTTCGTCCTTTCC
KLF1_gRNA_2F_b	**AGTCCAGGAG**GTTTTAGAGCTAGAAATAGCAAG
KLF1_gRNA2_SDMR	**GAGCGTACCT**CGGTGTTTCGTCCTTTCC
HBG_gRNA1_SDM_F	**ACAAGCCTGT**GTTTTAGAGCTAGAAATAGCAAG
HBG_gRNA1_SDM_R	**GATAGTAGCC**CGGTGTTTCGTCCTTTCC
HBG_gRNA2_SDM_F	**CTTCCTTTTA**GTTTTAGAGCTAGAAATAGCAAG
HBG_gRNA2_SDM_R	**CACCCTTCAG**CGGTGTTTCGTCCTTTCC
HBG_gRNA3_SDM_F	**CTAAGACTAT**GTTTTAGAGCTAGAAATAGCAAG
HBG_gRNA3_SDM_R	**AGTATCCAGT**CGGTGTTTCGTCCTTTCC
HBG_gRNA4_SDM_F	**GCCAACCCAT**GTTTTAGAGCTAGAAATAGCAAG
HBG_gRNA4_SDM_R	**CAGCCTTGCC**CGGTGTTTCGTCCTTTCC
HBG_gRNA5_SDM_F	**AGATAGTGTG**GTTTTAGAGCTAGAAATAGCAAG
HBG_gRNA5_SDM_R	**CAATGCAAAT**CGGTGTTTCGTCCTTTCC
Primers for sequencing over gRNA site	
CRISPR_SEQ1_F	GAGGGCCTATTTCCCATG
CRISPR_SEQ2_R	GTCGTTGGGCGGTCAG

Appendix 4

XRCC6 & LIG4 siRNA and rtPCR Primer Sequences. siRNA were purchased from Origene. rtPCR primers for Ligase IV, XRCC6 and KLF1 were designed using Primer3Plus[385], and rtPCR primers for Actin β, α-globin, β-globin and γ-globin are the same as those used by Mabaera *et al.* (2007)[550].

siRNA	Sequence	Ref Number
Ligase 4		
siLIG4-A	AGCUCAUACUAAGAAUGAAGUAATT	SR302689A
siLIG4-B	UCAAUAGACAAGUGUGAAUUACAAG	SR302689B
siLIG4-C	AGAAUGGCCUAUGGAAUUAAAGAAA	SR302689C
XRCC6		
siXRCC6-A	GCGCCAAAGUGAGCAGUAGCCAACA	SR301689A
siXRCC6-B	GUUCUAUGGUACCGAGAAAGACAAA	SR301689B
siXRCC6-C	CGAGGGCGAUGAAGAAGCAGAGGAA	SR301689C

Target	rtPCR Primers	Sequence
Ligase IV	*Lig4_cDNA_F*	CGAGCTTACCAGATGCCTTC
	Lig4_cDNA_R	TGTGGAACAGAGAAGCCAGA
	Lig4_cDNA_2F	CATGCAGGCTTGACAACATC
	Lig4_cDNA_2R	AGCCTGACCTGGAGAACAGA
XRCC6	*XRCC6_cDNA_F*	GGGACAAAAACGTTTCCAAG
	XRCC6_cDNA_R	CCAGGTTTCTTCAGGTGCAT
	XRCC6_cDNA_2F	CGGGAAACAAATGAACCAGT
	XRCC6_cDNA_2R	TGAAACCCATGAGCATCAAA
Actin β	*b_actin_cDNA_nF*	GTGGGGCGCCCCAGGCACCA
	b_actin_cDNA_nR	CTCCTTAATGTCACGCACGATTTC
α-globin	*HBA_cDNA_nF*	TGGGGTAAGGTCGGCGCGCA
	HBA_cDNA_nR	TGCACCGCAGGGGTGAACTC
β-globin	*HBB_cDNA_nF*	GGTGGTCTACCCTTGGACCC
	HBB_cDNA_nR	GATACTTGTGGGCCAGGGCA
γ-globin	*HBG_cDNA_nF*	GGGAGATGCCATAAAGC
	HBG_cDNA_nR	ATTGCCAAAACGGTCAC
KLF1	*KLF1_cDNA_EX1_F*	CTTCCCGGACACACAGGATG
	KLF1_cDNA_EX2_R	GGTCCTCAGACTTCACGTGG

Appendix 5

Primers used for Sanger sequencing over the site of the SNPs of interest in KLF1 and ASH1L.

Target	Primer	Sequence
KLF1	*KLF1_SNP_seqF*	GTTGCCCAGGCTACCTTC
	KLF1_SNP_seqR	GTGGGCTGGCTGGAATC
ASH1L	*ASH1L_SNP_seqF*	TCCTTTCTGTGAAGCCGATTTA
	ASH1L_SNP_seqR	AGTTCTCCAAGCTTATCCCTTG

Appendix 6

List of the 11 Fantom5 expression datasets used to generate the list of 'haematopoietically silent' genes.

Fantom5 data - Tissue Sources
Bone Marrow, Adult
CD34 cells differentiated to erythrocyte lineage, Biological Replicate 1
CD34 cells differentiated to erythrocyte lineage, Biological Replicate 2
CD34+ stem cells - adult bone marrow derived
Fetal Liver, Pool 1
Peripheral Blood Mononuclear Cells, Donor 1
Peripheral Blood Mononuclear Cells, Donor 2
Peripheral Blood Mononuclear Cells, Donor 3
Whole Blood (ribopure), donor 090309
Whole Blood (ribopure), donor 090325
Whole Blood (ribopure), donor 090612